proceedings from *computers & philosophy*, an international conference

i-C&P 2006

3-5 May 2006 - Laval, France

Scientific Officer:
Associate Professor Colin SCHMIDT

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED *(From – To)***
20-10-2008	Conference Proceedings	3 May 2006 - 5 May 2006

4. TITLE AND SUBTITLE

Computers & Philosophy, an International Conference

5a. CONTRACT NUMBER FA8655-06-M-5068

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Conference Committee

5d. PROJECT NUMBER

5d. TASK NUMBER

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Le Mans University
52, rue Docteurs Calmette et Guérin
Laval 53020
France

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
Unit 4515 BOX 14
APO AE 09421

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

CSP 06-5068

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Final Proceedings for Computers & Philosophy, an International Conference, 3 May 2006 - 5 May 2006

The increasing interaction between Philosophy and Computer Science over the past 40 years has lead to many position-taking stances in theories of mind, applied machine-embedded intelligence and cultural adaptations to the onslaught of robots in society. This volume constitutes a key contribution to the body of knowledge within or about the intersection of the two fields. Is there a proper answer to the question of whether machines can think? Contemporary thought on computers and Artificial Intelligence is not the exclusive aim of the project; the birth of original forms of machine intelligence can inform us about *potential* human beliefs and permissibility thresholds with regards to technology —i.e. are all communities equally-footed with respect to machines that speak? The texts in the present set of proceedings contain the full-length versions of the papers presented at the i-C&P conference. EOARD, along with local authorities, financed this scientific event bringing together specialists in the study of minds and machines from twenty-two countries. Nearly all the papers were presented in English. The texts are divided into four sections. The preliminary section contains *Keynote Addresses* including the "Paul Ricoeur Lecture" given by Francis Jacques (in French). *Section 1* contains the more technical papers, some of which are oriented towards studies in communication. *Section 2* contains solid work delving into slightly less scientific notions in order to bring the notions of the mind, cognitive science and values into discussion. The texts in *Section 3* take on the logical basis of mind, conceptual relations as well as interrogative techniques. The harmony of the authors work can be underlined thanks to a general focus on the notion of *thought*, hence the titles of the sections proposed.

15. SUBJECT TERMS
EOARD, Computer Science, Computers, Artificial Intelligence

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18, NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			JAMES LAWTON Ph. D.
UNCLAS	UNCLAS	UNCLAS	UL	734	**19b. TELEPHONE NUMBER** *(Include area code)* +44 (0)1895 616187

computers & philosophy

is organised by the
Computer Science Laboratory at
LE MANS UNIVERSITY

Held in cooperation with the
American Association for Artificial Intelligence*,*
European Office of Aerospace Research and Development
and IEEE France

For information on the i-C&P 2006 Laval programme
including a complete list of the papers presented,
abstracts, photos and poster
THE CONFERENCE WEB SITE MAY BE CONSULTED AT:

http://ateliers.iut-laval.univ-lemans.fr/i-CaP_2006//

We wish to thank the following for their contribution to the success of this conference:
European Office of Aerospace Research and Development, Air Force Office of Scientific Research, United States Air Force Research Laboratory
<http://www.london.af.mil>

SCIENTIFIC COMMITTEE

General Chair: Colin T. SCHMIDT, Communication, Philosophy & Cognition, Le Mans University, France

Local Organisations Chair: Xavier DUBOURG, Computer Science & Learning, Le Mans University & Director of the Laval Technological Institute, France

Honorary Chair: Francis JACQUES, Emeritus Professor of Philosophy, Sorbonne University, France

Varol AKMAN, Philosophy and Computer Science, Bilkent University, Turkey
Jean CAELEN, Cognition and Interaction, CNRS/Grenoble University, France
Raja CHATILA, Robotics, CNRS/Toulouse University, France
Nathalie COLINEAU, Language & Multi-modality, CSIRO, Australia
Roberto CORDESCHI, Computation & Communication, Salerno University, Italy
Liu GANG, Information & Philosophy, Inst. of Philosophy, Chinese Acad. of Soc. Sciences, China
Hans Peter DOMMEL, Computer Engineering, Santa Clara University, USA
Deborah G. JOHNSON, Technology and Ethics, University of Virginia, USA
Frédéric KAPLAN, Artificial Intelligence, SONY CSL - Paris
Nik KASABOV, Computer and Information Sciences, Auckland University, New Zealand
Oussama KHATIB, Robotics & Artificial Intelligence, Stanford University, USA
Boicho KOKINOV, Cognitive Science, New Bulgarian University, Sofia, Bulgaria
Felicitas KRAEMER, Philosophy & Intentionality, Bielefeld University, Germany
Jean LASSÈGUE, Philosophy, CNRS/Ecole Normale Supérieur Paris, France
Ping LI, Cognitive Science & Philosophy of Science, Sun Yat-sen University, China
Daniel LUZZATI, Linguistics, Le Mans University, France
M.C. MANES GALLO, Info. & Communication Sciences, Bordeaux University, France
Anne NICOLLE, Computer Science & Interdisciplinarity, CNRS/University of Caen, France
Teresa NUMERICO, Communication, Salerno University, Italy
James MOOR, Philosophy, Dartmouth College, USA
Bernard MOULIN, Computer Science, Laval University, Canada
Denis VERNANT, Logic & Philosophy, Grenoble University, France
Ming XIE, Robotics, Nanyang Technological University, Singapore

Dr. Colin Schmidt (Chair)

Computer Science Laboratory LIUM CNRS FRE 2730
Le Mans University
France
Tel +33 2 43 59 49 20 / +33 2 43 59 49 25
Fax: +33 2 43 59 49 28
E-mail: Colin.Schmidt@lium.univ-lemans.fr

We wish to thank Susan Fuller & Paul Losiewicz (EOARD) for their friendly cooperation.

introduction

The increasing interaction between Philosophy and Computer Science over the past 40 years has lead to many position-taking stances in theories of mind, applied machine-embedded intelligence and cultural adaptations to the onslaught of robots in society. This volume constitutes a key contribution to the body of knowledge within or about the intersection of the two fields. Is there a proper answer to the question of whether machines can think? Contemporary thought on computers and Artificial Intelligence is not the exclusive aim of the project; the birth of original forms of machine intelligence can inform us about *potential* human beliefs and permissibility thresholds with regards to technology — i.e. are all communities equally-footed with respect to machines that speak?

The notion of machines that have desires and beliefs, increase their own learning capabilities, develop bodily functions, play games with us, help us learn, help children or the ill to express themselves, care-give the elderly, etc. used to create heated debates. Or do they still do so? In view of these on-going investigations, comparative studies and forward-looking accounts are offered herein by a brilliant network of authors, as well as reports on innovative uses of knowledge found at the crossroads of philosophy and intelligent machinery sciences. Breaking news in computer science that pull the philosopher towards the computationalist point of view on mind are equally represented; and so are proposals that show the limits of representationalist theories. The main goal of holding the i-C&P 2006 Laval France was to spur on interdisciplinary dialogue between 80 engaging intellectuals.

The texts in the present set of proceedings contain the full-length versions of the papers presented at the i-C&P conference. EOARD, along with local authorities, financed this scientific event bringing together specialists in the study of minds and machines from twenty-two countries. Nearly all the papers were presented in English. The texts are divided into four sections. The preliminary section contains *Keynote Addresses* including the “Paul Ricoeur Lecture” given by Francis Jacques (in French). *Section 1* contains the more technical papers, some of which

are oriented towards studies in communication. *Section 2* contains solid work delving into slightly less scientific notions in order to bring the notions of the mind, cognitive science and values into discussion. The texts in *Section 3* take on the logical basis of mind, conceptual relations as well as interrogative techniques. The harmony of the authors work can be underlined thanks to a general focus on the notion of *thought*, hence the titles of the sections proposed.

The final papers received appear in this document. In this volume, the reader will find reference to further information on the papers presented, short abstracts and author contact information as well as further information on the sponsorship of the event.

Colin T. Schmidt
Le Mans University

Keynote Table of Contents

Session Speakers Table of Contents

SESSION 1

keynote speakers' contributions

QUELLE IMAGE DE LA PENSEE :
COMPUTATIO, COGNITIO, COGITATIO ?

FRANCIS JACQUES
Sorbonne University (Paris 3)

*Les plus 'profondes' questions : – Comment n'as-tu pas pensé à ceci ?
– Et toi, comment y as-tu pensé ? Paul Valéry*

La contribution que j'ai accepté de faire en hommage à mon maître disparu porte le titre : *The Paul Ricoeur's lecture Quelle image de la pensée : Cogitatio, computatio, cognitio.* Elle me donne l'occasion d'évoquer l'expertise de P. Ricoeur, d'en donner une réévaluation dans un domaine où le flambeau de l'anthropologie philosophique doit être tenu -- le domaine des sciences cognitives (SC).

Qu'on ne compte pas sur Ricoeur pour rester à l'extérieur d'une discussion dont il serait pratiquement exclu, à la manière de bien des philosophes traditionnels. D'autant que les scientifiques abordent des questions philosophiques traditionnelles, comme les universaux, l'analyticité, la mémoire, l'induction, la référence des mots aux choses. Mais dans un type de texte et selon un mode d'interrogation qui n'ont rien à voir avec ceux dont il a l'habitude. Surprise : certaines positions métaphysiques se trouvent réhabilitées au point qu'on a le sentiment que la philosophie est continuée par la science. Je dis tout de suite qu'il n'en est rien. Le mentalisme dualiste de Descartes qui fait de l'esprit une seconde substance donne lieu à un texte différent du mentalisme méthodologique de Chomsky qui considère l'esprit comme une sorte de puissance occulte sur quoi nous ne pouvons que spéculer indirectement à partir des performances linguistiques. De même pour la question des rapports de l'esprit et de la machine.

Paul Ricoeur a conduit un dialogue avec Jean-Pierre Changeux qui, vous le savez sans doute, avait été surnommé 'Mr. Cognition'. D'un côté, un représentant des SC, en fait des neurosciences, qui parlait au nom des

collectifs des théoriciens naturalistes du cerveau comme simple produit par l'évolution darwinienne. De l'autre, le champion de *l'acte de pensée* comme exercice de prise en responsabilité de soi pensant, par quelqu'un qui entendait ne laisser à personne d'autre la charge de répondre. Ce dialogue, publié aux éditions Odile Jacob, illustre les difficultés de la relation entre la phénoménologie dans la ligne de Husserl et les neurosciences[1].

Je voudrais revenir sur le ressort et le principe de ces difficultés, en rééchelonnant les trois termes qui partagent le rapport entre une philosophie de l'esprit et une science de l'esprit : *cogitatio*, *computatio*, *cognitio*.

Ces trois termes, chacun à sa manière, ont prétendu porter les chances d'une anthropologie rigoureuse : *cogitatio* pour le compte du dualisme, hier du corps et de l'âme, aujourd'hui du cerveau et de la pensée, avec les neurosciences et la phénoménologie. Le deuxième terme *computatio* penche vers un monisme ontologique, selon lequel c'est le cerveau qui pense en tant que structure organique capable de traiter l'information en calculant. On a reconnu l'objection que Thomas Hobbes opposa à Descartes. Le troisième terme *cognitio* convoque la philosophie d'aujourd'hui comme partie prenante dans le débat des sciences cognitives. A vrai dire, elle s'est trouvée concernée par la reprise explicite, d'aucuns diraient la relève de ses thèmes spécifiques, par naturalisation des concepts de la philosophie de l'esprit.

1. Phénoménologie et sciences cognitives : un dialogue bloqué

Par ce rééchelonnement, je prendrai la suite, mais sur une ligne différente, de la phénoménologie. Il apparaît en effet, à la lecture des quatre premiers chapitres du livre, que peu à peu le dialogue entre Changeux et Ricoeur sur '*ce qui nous fait penser*' s'est trouvé bloqué. Les questions des SC font directement écho à des thèmes philosophiques centraux, à des objets qui sont justement revendiqués en priorité par la phénoménologie. Elles ont produit en philosophie analytique un *cognitive turn*, qui réagit sur les autres branches (éthique, esthétique, théorie des normes). Trois dimensions de l'esprit sont visées par le projet naturaliste des SC : l'autonomie d'une science de l'intentionnalité ou phénoménologie ; les expériences qualitatives en première personne, le domaine de l'interprétation, des raisons et des normes.

1 J-P Changeux, P. Ricoeur, Ce qui nous fait penser. La nature et la règle, Paris Odile Jacob 1998.

D'un côté, le phénoménologue voulait restaurer une expérience qu'on peut résumer par 'le primat de la subjectivité'. Une expérience qui insiste en divers secteurs[2], alors qu'elle est dissimulée sous les objectivations scientifiques. D'un autre côté, le naturalisme cognitiviste qui entend biologiser le vécu, en réduisant à des systèmes mentaux du cerveau humain, les dimensions de l'expérience que la phénoménologie avait repérées et tenues pour irréductibles à des causalités factuelles : l'intentionnalité ou référence des représentations à l'objet, l'interaction communicationnelle, l'orientation de l'action vers un but, sa subordination à des normes ou à des valeurs.

Trois raisons

Je voudrais revenir sur ce blocage du dialogue et à ses présupposés latents, avant d'évoquer une alternative. D'où venait le blocage ?

1°) Première raison : le philosophe s'aperçoit qu'il n'a pas le monopole de la *cogitatio*. N'en déplaise à Heidegger, la science elle aussi 'pense'. Il n'y a pas d'un côté l'expérience pensante de la subjectivité avec son primat égologique, et de l'autre les causalités factuelles objectives, mais il y a bien deux interprétations concurrentes des données de l'expérience. Un pas de plus, ses objectivations passent du plan méthodologique au plan idéologique. Ce n'est pas seulement un scientifique que Ricoeur affrontait mais un philosophe naturaliste. L'option du naturalisme qui a actuellement le vent en poupe avec le retour d'une philosophie de la nature.

Je donne de cette première raison une lecture propre. La science elle aussi *pense* car elle constitue des faits en se faisant, i.e. en inventant ses problématiques, en déployant les catégories de son propre mode *d'interrogation*. On ne peut faire reculer cette option qu'en passant du plan herméneutique au plan érotétique d'une philosophie de l'interrogation. La moindre des choses est de recommander le *pluralisme* des interrogations. Une égologie peine à demeurer au foyer permanent de la phénoménologie, car les ressources fondationnelles de *l'ego* sont insuffisantes pour construire l'objectivité du sens. Ricoeur lui-même a toujours été partagé par une tension entre le pôle égologique à défendre contre les assauts successifs de l'objectivisme (Lacan, Lévi-Strauss, Greimas, l'institutionnalisme en matière de *Speech acts*). Et d'autre part, le refus de s'en tenir à une pure assomption de *l'ego*. De là son effort

2 Dans l'interprétation des symptômes névrotiques et des œuvres littéraires, aussi bien que dans l'interprétation des actes volontaires, dans les actes de parole de la promesse et du témoignage, que dans la saisie de la structure narrative de l'histoire ou de la fiction romanesque.

pour dialoguer avec la philosophie analytique. Il devait opter pour une constitution langagière ou narrative du sujet, cependant que l'expérience vécue, de moins en moins convoquée, reculait à l'horizon de l'analyse du discours.

2°) La deuxième raison du blocage devient évidente. La pratique du dialogue s'est refermée comme un piège sur Paul Ricoeur. Que devient le dialogisme dans une controverse méta-théorique où chacun réactive sa propre position. Alors que Ricoeur s'était fait une règle d'accueillir la parole d'autrui, son interlocuteur était bien décidé à distribuer les cartes à son avantage en s'attribuant par principe celles du gagnant. Une ambiguïté : Jean-Pierre Changeux, au nom de la science cognitive mais d'obédience naturaliste, entendait d'emblée et jusqu'au bout 'partir à la conquête de l'esprit'. Dès lors, comme le dit Jean-Luc Petit : dans l'esprit du biologiste naturaliste, le phénoménologue peut bien revendiquer l'exercice de la réflexion sur l'esprit humain 'au mieux, il redira en termes approximatifs ce dont le biologiste est seul à penser selon un savoir empiriquement fondé'[3]. Dans ces conditions, le parti pris dialogique, maltraité dans la pratique, n'était plus d'aucun secours.

3°) Après avoir décelé les présupposés cachés d'une discussion, il faut répartir les *tâches* des protagonistes de la discussion, suivant leur mode d'interroger, problématisante chez le biologiste, radicale et principielle chez le philosophe.

La première tâche du philosophe est la *vigilance épistémologique* à l'égard des homologations disciplinaires. Les SC se veulent davantage qu'une simple 'coalition opportuniste de disciplines en mal de subvention', elles se donnent comme une conjonction synergique à visée authentiquement interdisciplinaire, fédérée par une problématique commune : le traitement de l'information, dans le cadre d'un paradigme computationnel. Tel est l'horizon. Il peut servir de repère critique.

Or, la mise en œuvre de cette interdisciplinarité se heurte à *plusieurs amphibologies*. A propos du programme de recherche sur l'inscription neuronale des représentations sociales et même des représentations éthiques de soi et des autres, comment s'orientait la critique philosophique de Paul Ricoeur ? Il essayait de *déminer* par l'absurde les problèmes *d'interface* entre une science qui a son centre de gravité dans la biologie neurologique et une science des comportements sociaux. C'est un fait que l'expression 'représentation sociale' figure dans le lexique des neurosciences au titre d'objet mental, et aussi dans le lexique de l'anthropologie culturelle. Il s'agit tantôt d'image interne élaborée par le cerveau en réponse aux informations de l'environnement, tantôt de

3 J-L Petit, 'Sur la parole de Ricoeur : « le cerveau ne pense pas'. Je pense. », Revue d'histoire et de philosophie religieuse, Strasbourg janvier-mars 2006.

croyances formulées en propositions de forme ‘x croit que’, tantôt encore de formations sociales à finalité de communication :

> Il y a intérêt à ce que chaque discipline reste maîtresse de son ordre du jour, soit l’organisation neuronale soit les formes sociales de communication[4].

Cette objection mériterait d’être élargie et sondée dans son fondement philosophique (infra 3ème partie). Il en va des procédures acceptées comme valables pour la modélisation des hypothèses et leur confirmation. Seule la discussion *interdisciplinaire* est habilitée à convenir des zones de recouvrement qui permettent de confronter les résultats. Aucune discipline isolée, la biologie cognitive pas plus qu’une autre, n’est autorisée à résoudre les problèmes d’interface à l’intérieur de son propre champ.

De quel poids pèsent les contraintes exercées les unes contre les autres ? Est-il légitime de mêler des usages aussi différents que la trace cérébrale, la trace psychique et la trace culturelle ou inscription ? Celle-ci propose une métaphore sémiologique qui ne se réduit ni aux bases neuronales de la conscience ni aux traces vécues d’un événement marquant. Avec ce déplacement, se pose un problème d’accueil dans le champ des sciences neuronales, un autre dans le champ des sciences humaines.

Normalement, le dialogue du biologiste naturaliste et du philosophe chrétien aurait dû déboucher sur un discours commun. Seulement, la naturalisation des concepts de la philosophie de l’esprit par le biologiste du cerveau se heurtait à une ambiguïté plus marquée encore. Les deux protagonistes sont loin de pouvoir construire un discours explicatif homogène pour rendre compte des niveaux de complexité en termes de mécanismes moléculaires, chimiques ou électriques. Le terme même de *mécanisme* ne garde pas le même sens quand on saute d’un niveau de description à un autre, du mécanisme au plan mental au mécanisme au plan neuronal.

Que dire de la proposition de ‘naturaliser l’intentionnalité’ en la comprenant comme le niveau de représentation le plus élevé, alors que pour les philosophes il est précisément impossible de refermer la notion *d’intentionnalité* sur celle de représentation ? Comme le reconnaît J.P.Changeux avec une certaine désinvolture épistémologique, ‘les jeux de langage sur le mot représentation ne le concernent pas’.[5] Il est tout près de franchir les lignes de fracture entre disciplines :

> On s’expose certes aux dangers de l’interprétation illégitime, mais on prend aussi le

---

4 J-P Changeux, P. Ricoeur, op cit p. 169.

5 J.P. Changeux et P. Ricoeur op. cit 171 et 175.

> risque de faire des découvertes importantes'[6].

Il ne revient pas au même de traiter de la *fonction référentielle* du langage ou de la propriété que possèdent les représentations dans le cerveau d'être au sujet des objets qui les causent. Il ne revient pas au même de traiter de la *représentation* comme l'acte de se représenter un objet ou de la traiter comme d'un objet mental dans le cerveau.

L'épistémologue relève que la modélisation de l'espace conscient est présentée comme un fait acquis et il note avec une certaine malice que la construction du modèle neuronal est significativement plus avancée que sa vérification expérimentale. D'où vient cette avance et est-elle légitimement obtenue ? :

> Cette avance, vous la devez aux progrès faits dans des disciplines qui ne doivent rien aux sciences neuronales. Après coup, vous intégrez leur résultat en vous efforçant de rester en accord avec vos prémisses de base. Lesquelles limitent la portée des analyses empruntées[7].

Le philosophe comme l'épistémologue se trouve devant la tâche de distinguer sous l'uniformité apparente du domaine des SC, la diversité réelle des approches dont chacune engendre son propre discours et sa propre ontologie. Même s'il s'agit de la même chose, 'le traitement de l'information' – ce qui fonde les SC comme interdiscipline, sous une problématique procédurale commune – ce *traitement* ne se laisse pas traduire de discipline en discipline sans perdre de sa spécificité. Le terme *d'information* n'a pas le même sens pour la sémantique de la référence, l'écologie du rapport du vivant avec son milieu, l'éthologie du comportement animal, la neurologie des systèmes cérébraux, et la théorie mathématique de l'information. S'agissant de *l'imagerie cérébrale* qui est censée nous faire voir les états mentaux, il n'est pas certain non plus que les images obtenues par des techniques différentes soient comparables. Il n'est pas d'image en soi dans le cerveau. Toute image dépend d'une démarche d'aspectualisation du réel chez qui l'interroge et le catégorise.

Nous versons ici dans la seconde tâche : caractériser le discours des protagonistes de la discussion en fonction de leur *mode d'interroger*. Il convient de souligner, en face de l'interrogation formelle et problématisante de la science, l'existence de l'interrogation philosophique qui est informelle et radicale, ainsi que de l'interrogation théologique qui procède à l'élucidation conceptuelle du mystère révélé et même de l'interrogation poétique qui produit les belles énigmes de la condition humaine. Résumant de longues analyses, je préférerai au sujet langagier ou narratif de Ricoeur, trop égologique, mettre en avant le statut du *nous* inter-rogeant sur lequel *l'ego* a toujours à se ressaisir. Qu'on y songe : le

---

6 Ibid.

7 Ibid 147.

seul lieu sémantique où les personnes se *rencontrent* effectivement, c'est l'inter-rogation. L'interrogeable déborde telle ou telle figure de l'interrogé. Ou, pour le dire de manière classique, l'être ne se réduit pas à la nature et la nature ne se *réduit* pas aux décisions naturalistes.

Du naturalisme au réductionnisme

Ils ont partie liée. Le naturalisme prétend assurer la relève de la philosophie : emprunter à la science de quoi la continuer par d'autres moyens, en effaçant la démarcation entre l'homme et les autres vivants. C'est la doctrine suivant laquelle la vie morale prolonge la vie biologique et l'idéal moral n'est que l'expression des instincts constitutifs du vouloir-vivre. Primauté morale de la perpétuation de la vie, homogénéité des fins humaines et des fins animales. Le naturalisme comme méthode de déterminisme, très bien. Mais comme doctrine exclusive ? Ne convient-il pas d'imaginer une nouvelle philosophie de la nature consciente du fait que 'l'objet naturel' n'est pas la réalité tout entière mais n'en est qu'un *aspect* ? D'autres approches seraient bien venues pour en corriger en permanence l'abstraction réductiviste.

La vigilance du philosophe doit s'exercer aussi à l'égard du réductionnisme Tout art représentatif utilise un système de *réduction* : la sculpture réduit la forme à deux dimensions ; la sculpture réduit le mouvement à l'immobilité. Construire, c'est simplifier, avant de compliquer de manière réglée. Il faut bien réduire méthodologiquement la diversité du réel pour objectiver ou modéliser. Les procédés ne manquent pas : confondre condition nécessaire et condition suffisante. Faire jouer le tiers exclu selon une analogie dominante. Fermer l'analyse prématurément, en rendant réversible l'analyse et la synthèse. Descartes réduit l'être à la nature, celle-ci à l'étendue intelligible, l'esprit à la pensée, à la *cogitatio*. Procédés logiquement peu défendables. C'est attenter dogmatiquement à la *différence* au mépris des niveaux d'organisation et de leurs lois propres. On traite toute la physico-chimie par le seul emploi de la mécanique des solides, les êtres vivants par le seul emploi de la physico-chimie etc.

Il est certes nécessaire de réduire pour connaître ou agir. Seulement, la réduction a ses limites. En général, nous tirons vers le bas ou vers le simple. Nouvel indien Jivago, avec une gaîté de plus ou moins bon aloi, nous voilà réducteur de têtes. La question se pose quand nous réduisons la foi à la croyance, la transcendance à l'immanence, l'inconditionné au conditionné, le sujet transcendantal au sujet empirique, *l'homo* sapiens à *l'homo faber*, l'icône à l'idole, et sans doute aussi *homo* à ses caractéristiques biologiques. L'homme qui construit ne s'inquiète pas de toutes les

qualités de l'être ou de la chose qu'il modifie, mais seulement de quelques-unes, afin de lui donner prise sur eux. Il tend à faire abstraction du singulier. Il retient du réel un modèle permanent en problématisant son objectivité. Il s'attache seulement à des conditions claires et distinctes qu'il peut manipuler. Lui suffisent celles qui répondent à son but et qui peuvent être satisfaites simultanément.

C'est juste l'inverse pour l'homme qui aime. Aimer tend à dilater l'objet à la mesure de l'être. A rebours de la *réduction*, il augmente le nombre des qualités perçues ou désirées. Il pose la permanence de la personne singulière avec le sentiment de lui rendre ce qui lui est dû. La saisie de celui qui aime ne relève plus du problème mais du mystère, au croisement de l'inaccessible et de la lumière. Stendhal parlait de 'cristallisation' ; Valéry s'étonnait qu'aimer rende 'inventif, prévenant', 'inépuisable'.

Pas la même topique. Contre le réductionnisme vaut une approche de *sursomption* qui ne nie pas ce qu'elle dépasse, mais en manifeste l'insuffisance. Au lieu de réduire le supérieur à l'inférieur, elle montre comment l'inférieur s'achève, s'accomplit et se justifie dans le supérieur. Le surgissement d'un nouveau niveau de sens reprend rétrospectivement le sens d'un dépassement-reprise du niveau inférieur. Variété de topique transcendantale, elle part en guerre contre le positivisme. A l'idée de *réduire*, j'opposerais l'idée de *reconduire*. Pascal voulait comprendre par le supérieur, à partir d'un point haut, site projectif qui éclaire l'inférieur et le situe. Je tiens, quant à moi, à la convergence, à la complémentarité des ordres du pensable, où le conflit de compétence est surmonté par la grâce des reprises consenties selon l'analogie interrogative.

A rebours de la totalisation, le réductionnisme est solidaire d'une approche analytique qui prétend appréhender un système en le décomposant en éléments suffisamment simples. Le risque est d'ignorer les interactions à l'origine des propriétés *émergentes*. L'approche réductiviste s'intéresse à la nature des interactions, alors que l'approche *systémique* s'intéresse aux effets de l'interaction. Holistique, elle s'applique aux systèmes trop complexes pour être décomposés. Le risque de réduire ? Ecraser des distinctions essentielles : entre niveaux d'organisation, concepts, catégories. Une dimension de l'être (ou de la connaissance) est interprétée comme identique à l'être (ou à la connaissance). Elle réduit à l'homogène, avec un alibi explicatif. Par omission ou oblitération d'une fracture ou d'une *émergence*. E.g. la pensée n'est que le cerveau ; l'œuvre n'est qu'un travail ou qu'un jeu.

D'un point de vue *logique*, l'opération consiste à assimiler une classe d'objets à une autre. Elle entraîne deux sortes de réduction : conceptuelle, quand un terme est définissable par d'autres qui désignent des entités observables. Propositionnelle, quand on transforme des énoncés pour les amener à respecter un modèle d'intelligibilité. A condition que les valeurs de vérité soient préservées.

D'un point de vue *épistémologique*, la réduction consiste à subsumer une théorie sous une autre, au moyen de règles de correspondance. On tente de définir les énoncés d'une discipline (e.g. psychologie) par des concepts d'une discipline plus opératoire (e.g. neurophysiologie). On s'efforce de dériver les lois propres à une science des lois propres à une autre. Nous avons deux propositions. Le réducteur pose leur équivalence en arguant que la vérification de la seconde vérifie la première.

D'un point de vue *linguistique*, les terribles formules réductives en *ne…que* sont d'autant plus péremptoires qu'elles se veulent équationnelles. Je dis : la pensée n'est que le fonctionnement du cerveau. C'est plus qu'une obsession grammaticale ou qu'une forme de révérence envers certaines tournures de langage.

Comme Wittgenstein le relève dans une conversation avec Waismann (1931), on peut donner de la réduction soit une version non dogmatique en posant simplement une règle de grammaire selon laquelle 'la première proposition doit suivre de la seconde'. Au lieu de parler du sens, on reste à l'intérieur de la grammaire. Soit une version dogmatique en disant qu'une proposition est vérifiée de deux manières différentes, et que dans les deux cas, elle a même sens. En fait, une forme de description est préférée et imposée. Le réducteur renonce à voir la description initiale comme celle d'un phénomène premier. La réduction apparaît comme une *préférence* en faveur d'une forme de description, non pas une obsession grammaticale ou un préjugé. Elle incline à appréhender les choses comme plus simples qu'elles ne sont en réalité.

Pour une part, la réduction fonctionne comme *interprétation*. La bonne interprétation est celle qui permet à l'interprète de se retrouver chez lui dans un certain langage, au point de ne plus envisager de dire et de voir les choses autrement. L'interprète nous conduit d'une forme d'expression à une autre. Le réductionniste fait un pas de plus : il ferme la possibilité de revenir au langage initial, désormais disqualifié.

Un réductionnisme *modéré* n'est pas sans pouvoir heuristique, dans la mesure où il est animé par un désir réel et respectueux d'unification théorique. Ce programme est loin d'être rempli. Trois raisons : d'abord, la science moderne par sa théorie du chaos ou la théorie du flou quantique introduit structurellement des éléments de non-déterminisme. L'idée d'une matière objective, constituée de parties séparables, connaissables de façon maîtrisée disparaît. Ensuite, les équations qui régissent l'objet observé dépendent du processus d'observation et celui-ci de l'existence d'un observateur. Même si le déterminisme absolu était vrai, il serait un postulat non vérifiable. La réduction de la conscience à la matière n'est plus fondée en rigueur de termes.

Chacun est prêt à accepter le parti pris réductionniste pour son compte, quand il lui sert. Quitte à le dénoncer chez les autres. En quoi le terme est polémique. On agite le tocsin. La question devient volontiers véhémente quand le théorème réductionniste - *x* n'est rien d'autre que *y* - a une portée *dogmatique*.

-- *positiviste* : l'expérience n'est que l'observable. Notre siècle est friand de simplifications qui se révèlent intenables. La plus récente est de réduire l'humain ou le primate à une simple machine à traiter de l'information. C'est la même époque, la nôtre qui conduit des recherches sur les singes parlants et sur l'intelligence artificielle.

-- *matérialiste*. On suppose qu'un état mental correspond à un état physique cérébral et à un seul et que cette correspondance entraîne la réduction de l'un des facteurs à l'autre. A noter qu'on pourrait accepter qu'un état mental n'existe pas sans traduction physique corrélée, sans pour autant accepter un déterminisme matérialiste.

-- *techniciste* : puisque c'est techniquement possible, c'est aussi moralement juste. L'art et la mathématique, le rite et la prière, les jeux de langage et la mise en question philosophique sont également des activités par lesquelles l'homme devient agent et patient de son agir. Ils lui sont tout aussi coextensifs.

-- *naturaliste* : *Homo* réduit à ses caractéristiques biologiques. L'éthique est réduite à l'analyse des conditions naturelles de l'action. La représentation des valeurs résulterait de la contrainte exercée par les besoins élémentaires ; les œuvres de la culture de la sublimation de pulsions sexuelles ou des conditions socio-économiques.

-- *érotétique* : le refus d'interroger encore et autrement entraîne la réduction abusive. Exemple : le statut de *l'embryon*. Les possibilités interrogatives de la pensée sont réduites à l'une d'entre elles ; l'assimilation abusive de deux modes d'interrogation est facilitée : l'énigme est réduite au problème ou le mystère à l'énigme. L'opération élémentaire est la reprise non critique d'une question. La liste n'est pas close.

Mieux vaut déconstruire le théorème réductionniste : *x* n'est rien d'autre que *y*. Il repose sur la légitimation du plan de pensée où se meut la pensée de *y*. La réduction est de fait, le réductionnisme de droit. Il nous enjoint de nous incarcérer dans une réalité unidimensionnelle. La perte sémantique est sévère. Un coup de force prétend autoriser le *passage* e.g. de la transcendance à la sainteté, de l'être-au-monde à la création. Faire la part de ce qu'on exclut, de ce qu'on perd et de ce qu'on explique vraiment.

On doit à D. Davidson un effort pour distinguer plusieurs théories de la relation entre événements mentaux et événements physiques. Certaines théories affirment l'existence de lois psychophysiques, d'autres la nient. Certaines théories affirment l'identité des événements mentaux et physiques, d'autres la nient. Il faut desserrer l'étau du réductionnisme : à côté du monisme nomologique (il y a des lois et les événements corrélés sont identiques), il faut maintenir la possibilité d'un dualisme monologique en termes de parallélisme ou d'interactionnisme). Davidson développant pour son compte un monisme anomal.[8] Le cartésianisme étant représentatif d'un dualisme anomal (absence de corrélation et dualisme ontologique).

Le réductionnisme *extrême* nous invite à faire prévaloir dans ses reprises la catégorie de condition (Eric Weil), pour atteindre quelque chose comme le *real meaning*. Si les sciences positives abordent le vivant, elles ne verront que des cellules et des molécules ; si elles abordent la pensée, que des produits de réseaux de neurones activés ; si c'est l'amour, que des processus hormonaux. Si c'est la foi, qu'un système de croyances, consolidé par l'institution. Tôt ou tard, le réductionnisme extrême se traduit par un mépris de l'homme.

La science ne pourrait expliquer l'articulation action-liberté-éthique qu'en intensifiant le réductionnisme. Ce que n'hésite pas à faire le darwinisme social. Mais le programme réductionniste présuppose une homogénéité des objets de connaissance qu'il est difficile d'accorder aujourd'hui. Le réductionnisme extrême qu'on rencontre parfois dans les neurosciences implique davantage que la simple correspondance d'un état mental à un état cérébral. Une telle correspondance pourrait exister sans que l'un des aspects soit réduit à l'autre. Encore faudrait-il qu'on puisse recréer un état mental par simple déclenchement du support physique. Or, il arrive que l'état physique visé ne puisse être créé à volonté ou qu'il se produise sans l'état mental. La théorie exigerait que l'état du cerveau et les idées correspondantes fussent expliquées et prévues.

2. De la phénoménologie à l'érotétique : quelle image de la pensée ?

En outre, le réductionnisme n'est jamais qu'une option *érotétique*, une certaine façon d'interroger la réalité. Nous retrouvons cette idée pour nous décisive. Il peut être en effet *méthodologique* quand on refuse provisoirement de s'intéresser à certaines questions qu'on ne sait pas

8 D.Davidson, 'Mental Events', Essays on Actions and Events, Oxford University Press 1980, tr.fr., PUF 1993.

encore traiter. Il peut aussi être *doctrinal* quand il transforme le statut ontologique des données brutes. L'homologie réductionniste est d'autant plus boiteuse que les types de questionnement diffèrent. Elle devient meurtrière quand la capture idéologique opère par glissement de terrain homicide.

Mon option propre consiste à recommander un pluralisme des modes d'interroger qui définissent la pensée comme *cogitatio* et assure la plénitude de son exercice.

Cogitatio et computatio

La *Cogitatio* ou pensée équivalait pour Descartes à la conscience dans toutes ses dimensions. Le moi pensant, dit-il dans la troisième Méditation, est 'une chose qui pense, c'est à dire qui doute, qui affirme, qui nie, qui connaît peu de choses, qui en ignore beaucoup, qui aime, qui hait, qui veut, qui ne veut pas, qui imagine aussi et qui sent'. Husserl la définit dans *Logique formelle et logique transcendantale*, en tant qu'elle 's'accomplit dans le langage' et qu'elle est liée 'absolument au discours'.

Dans les *Objections* (iii, 4) publiées en 1641 avec la première édition des *Méditations* de Descartes, Hobbes déclare :

> Que dirons-nous maintenant, si le raisonnement n'est pas autre chose que le fait d'unifier et enchaîner par le mot 'est' des noms ou des désignations ? La conséquence sera que la raison ne nous livre aucune conclusion sur la nature des choses mais seulement sur les termes qui les désignent, à savoir l'existence d'une convention, faite arbitrairement à propos de leur signification, selon laquelle nous joignons ces noms ensemble.

Le *De Corpore* de 1655 porte le sous-titre séparé '*Computatio sive Logica*', qui doit introduire à tout son système. Thomas Hobbes y est plus explicite : 'les premières vérités furent constituées arbitrairement par ceux qui, les premiers mirent des noms sur les choses, ou les reçurent de l'imposition faite par d'autres. Car il est vrai e.g. que 'l'homme est une créature vivante' ; mais c'est pour la raison qu'il a plu à des hommes d'imposer à la fois ces noms sur la même chose[9].

> Les axiomes d'Euclide comme 'le tout est plus grand que ma partie' ne sont pas des principes de démonstration, dit-il c'est-à-dire des vérités à accepter sans preuves, mais des propositions qui sont démontrables elles-mêmes à partir de définitions (chap. §25).
> Elles doivent donc être distinguées des lois de la physique, qui ne sont pas

9 La traduction anglaise a été publiée en 1656 : Elements of Philosophy Concerning Body ,chap. 3, § 8,9.

constituées par des définitions arbitraires (chap. 26, §1).

Tel est le début de la théorie conventionnaliste de la vérité nécessaire. Dans le développement de la philosophie de Hobbes, elle est liée à la doctrine selon laquelle la pensée n'est qu'une manipulation de signes, et le titre *Computatio sive logica* suggère que la pensée raisonnante peut être réduite à une espèce de calcul. Cette suggestion a été reprise par Leibniz. Cependant que Berkeley y voyait une conséquence naturelle de sa propre philosophie empiriste. Il écrit dans son Notebook :

> La raison pour laquelle nous pouvons démontrer si aisément sur des signes est qu'ils sont parfaitement arbitraires et en notre pouvoir. Que deviennent les vérités éternelles ? Elles disparaissent… Retirer les signes de l'arithmétique et de l'algèbre, que reste-t-il ? Ce sont des sciences purement verbales sans autre utilité que pratique dans la société des hommes. Elles ne contiennent ni connaissance spéculative ni comparaison d'idées.

Ces suggestions n'ont été publiées qu'à la fin du $19^{\text{ème}}$ siècle et l'alliance étroite du conventionalisme et de l'empirisme n'a été établie qu'à notre époque. C'est seulement dans les années 30 qu'on parviendra à avoir une idée claire des propriétés qui constituent la *calculabilité*, en définissant la notion d'algorithme, en établissant la théorie des fonctions récursives et la théorie des machines de Turing, et surtout en unifiant ces résultats par la thèse de Church. Le calcul, c'est toute procédure qui permet d'obtenir de façon quasi automatique le résultat de n'importe quelle opération. Leibniz cherchait à construire un système logique qui fonctionne comme un calcul.

Le thème selon lequel le calcul est un modèle pour le fonctionnement de la pensée se rencontre chez Condillac et Boole. La nature d'un calcul ne dépend pas de son objet arithmétique, seulement une fonction n'est calculable que si elle correspond à une fonction arithmétique. Les algorithmes sont des recettes destinées à résoudre certains problèmes qui comportent la répétition d'une opération. Ce sont des suites finies d'opérations déterminées au préalable, et dont l'exécution sur une machine est prescrite univoquement par un texte fini.

L'émergence des SC est liée à la simulation sur ordinateur pour la modélisation. Le cadre théorique est le paradigme computationnel du traitement de l'information. Le débat porte sur le rôle des nombres, la nature des problèmes[10], la propriété métalogique de décision[11], bref sur la

10 Etant donné un nombre n décider si n est premier. On peut essayer de diviser n par 2, 3,…, n-1. Il y a des méthodes plus rapides. La différence entre le problème numérique ou non numérique n'est pas importante pour la nature des algorithmes. Etant donné un mot français m trouver la page où m est écrit. On ouvre le dictionnaire au hasard. On compare m avec le premier mot de la page. Suivant le résultat, on regarde les pages qui suivent ou qui précèdent.

limite de la *computatio* autour de la calculabilité au sens strict qui apparaît clairement. La science calcule, mais avant de calculer elle théorise et avant de théoriser elle problématise, en prenant l'initiative de construire des modèles rapportés aux observables de son champ.

Cogitatio et cognitio…

Le 'et' n'a pas forcément une fonction disjonctive. La réflexion de la philosophie de l'esprit se mêle au développement des SC. Elles prennent la *cogitatio* elle-même comme *objet* d'étude. Elles cherchent à développer un formalisme qui nous permette de décrire le savoir. Nous tentons de découvrir les 'atomes' et les 'particules' qui le constituent ainsi que les forces qui agissent sur lui[12]. Le problème est la représentation du savoir. Elles situent la *cogitatio* au sein de configurations dont elle est autant le produit que la productrice. Loin que le savoir objectif sur l'homme soit un produit du sujet humain, il est un doute émis à l'encontre des produits de ce sujet.

L'anthropologie philosophique ouvre la voie quand on se demande comment l'homme dispose d'une possibilité supplémentaire par rapport aux systèmes ouverts sur l'environnement de type biologique : celle de transmettre des normes. De son côté, la philosophie analytique qui s'attache à clarifier la description de *l'action* est préalable aux questions normatives, en assurant la spécification des actions volontaires par rapport au mouvement naturel, mieux : elle précède toute simulation de l'action par les SC. Faut-il expliquer l'action par des raisons ou par des causes demandait Wittgenstein dans *Le Cahier bleu et le Cahier brun* ? Une raison cause une action comme telle si elle ne cause un effet *que sous* une certaine description. Comment du mental peut-il causer du physique, si l'on rejette l'idée d'une réduction des raisons aux causes ? Les propriétés mentales dépendent des propriétés physiques au sens de la survenance[13].

---

11 Ce terme relève de la méta-théorie d'un système et concerne un problème. Un problème est décidable si nous disposons d'une procédure conduisant après un nombre fini d'étapes à répondre par oui ou par non. Si la procédure est effective, on exprime son contenu par la notion d'algorithme. On peut ainsi résoudre uniformément par un calcul tout problème d'un certain type. Si une théorie est décidable (oui pour l'algèbre et la géométrie élémentaire selon Tarski) et si nous connaissons une procédure de décision, nous sommes en mesure de résoudre sans effort d'invention tous les problèmes formulables dans le langage de la théorie (Non, pour le dernier théorème de Fermat. Oui, quand on veut déterminer si une équation algébrique a une racine rationnelle).

12 Winograd, Artificial intelligence and Langage Comprehension, 1976.

13 P.Engel propose de traduire le terme anglais 'supervenient' pour désigner une relation intermédiaire entre l'émergence et la simple covariance. Il s'agit d'une détermination sans réduction du mental vis à vis du physique. Voir sa présentation de la trad.fr. de

A l'inverse, les interrogations que suscitent les SC sont reprises dans les débats de la philosophie analytique de l'esprit et contribuent à les reformuler. Empruntons au programme de recherche lancé par Newell et Simon un exemple de synergie entre la philosophie de l'esprit et l'intelligence artificielle. A différentes phases de son déroulement sont soulevées les questions philosophiques de savoir si :

1°) l'esprit est en fait un système formel. Comprendre consistant à manipuler les représentations analysables en éléments primaires reliés par des relations syntaxiques.
2°) la faculté de comprendre nos émotions et nos pratiques sociales est réductible à un ensemble de croyances.
3°) la compréhension courante consiste en des savoir-faire quotidiens (G.Ryle : *knowing how*), indépendamment des 'savoir-cela' (*knowing that*).
4°) la compréhension courante (de la langue naturelle, de la parole prolongée, d'une situation en cours de changement) n'est que le produit d'inférences sur des faits assertés que nous connaissons déjà, explicitables indépendamment du contexte[14].

A supposer que ce soit le cas, ces difficultés philosophiques créent des *problèmes* pour l'I.A. qui ne sont pas résolus et peut-être insolubles :
1°) emmagasiner dans l'ordinateur la masse des croyances qui constituent la forme de vie humaine et savoir quel langage utiliser à cet effet.
2°) disposer de règles pour extraire les faits indépendants du contexte mais pertinents dans des contextes particuliers.
3°) trouver les critères de pertinence tout en clôturant la liste mal définie des types de contexte.
4°) fixer dans l'ordinateur le savoir-faire humain hors contexte. C'est le *frame problem* qui revient à représenter que, dans l'action en cours, certains faits changent mais pas tous et que seul un petit nombre de ces changements est pertinent.

Il est de fait que la *cogitatio* naturelle résout ces difficultés : il faudrait que les sciences de la *cognition* parviennent à modéliser les problèmes correspondants. L'analogie avec les ordinateurs permet d'essayer de comprendre le fonctionnement de l'esprit humain. Il évite autant le dualisme cartésien que le réductionnisme. Les spécialistes en information utilisent le critère comportemental pour l'intelligence. Ils construisent des ordinateurs tels qu'ils font preuve de comportements qui seraient qualifiés d''intelligents' s'ils avaient lieu chez des êtres humains.

D.Davidson, Essays on Actions on Events 1963-1978, Actions et événements, Paris, PUF 1993, p25 et note 1, p 286-287. Pour des analyses autrement orientées voir D.Vernant, Du Discours à l'action. Etudes pragmatiques, 1997.

14 F.Jacques, Ecrits anthropologiques, texte 7 et texte 2, Paris L'Harmattan 2000.

Aucune hypothèse naturaliste n'est rendue nécessaire par la stimulation cognitive.

C'est aussi le cas pour une théorie qui a pour allié le paradigme computationnel développé par les SC : la théorie fonctionnaliste des rapports entre le corps et l'esprit[15]. Car le niveau de description fonctionnelle ne coïncide pas avec le niveau de description physique. Un état mental ne diffère pas d'un état physique puisqu'il peut exercer une action causale sur lui mais il peut être réalisé physiquement de plusieurs façons. C'est l'idée que les états mentaux sont des états fonctionnels et que les relations entre eux sont causales. Mais elles aussi peuvent se réaliser de manière multiple. On reste en deçà du programme originel de Alan Newwell et Herbert Simon : l'hypothèse d'un système de symboles physiques postule que l'intelligence humaine et l'ordinateur digital sont deux exemples d'un type unique de mécanisme physique qui génère des comportements intelligents en manipulant des symboles avec des règles formelles [16].

Descartes dit : l'âme pense toujours. Pourquoi toute cette mnémotechnie sinon pour combler les lacunes de l'oubli et les saillies du Malin génie qui habitent l'intervalle, pour prolonger l'écho du présent momentané dans la pensée ? Mais la pensée n'est pas comme un coffret à idées dont tu tires celle dont tu as besoin. Ni trésor d'idées représentatives ni synthèse du jugement vrai, mais une image *processuelle* dans la dynamique pensante de la *cogitatio*.

Ce ne sera pas davantage l'image kantienne de la pensée. Que mettre en face de l'image dogmatique de la pensée, où penser, c'est affirmer ? Qu'il s'agisse de science, de poésie, de religion, la pensée se corrige sans cesse en rejetant les réponses et en reformulant les questions. Il y a là plus qu'une condition d'exercice de la pensée. Je propose d'habiter autrement les lieux obligés du concept, du jugement et du discours : c'est à l'interrogation de préparer dynamiquement la synthèse prédicative. *Energeia* plutôt *qu'ergon*. Le nouveau penser, une fois pensé sans concession, *c'est* l'interroger.

Au vrai, tout ne commence pas avec le jugement. La synthèse du jugement doit se faire. Il y a une dynamique du jugement qui se cherche dans et par l'interrogation. On ne suspend pas l'interrogation mais seulement l'assertion affirmative ou négative qui exprime le jugement. Penser, c'est construire une *interrogation*. Il s'ensuit : 1°) il vaut mieux enseigner comment penser que ce qu'on doit penser. 2°) Penser la

---

15 P.Jacob, 'Le problème du rapport du corps et de l'esprit aujourd'hui', essai sur les forces et les faiblesses du fonctionnalisme', in D.Andler, éd., Introduction aux SC, Paris, Gallimard, 'folio essais', 1992.

16 'Computer Science as Empirical Enquiry : Symbols and Search', A.Newell et H.Simon. Communications of the ACM, vol 19 n°3 mars 1976, 116.

pensée, c'est interroger l'interrogation, sa puissance et son envergure en général. Sa distribution en modes stables que l'on peut comparer, associer, recouper. Telle que je la pratique, la philosophie de l'esprit est attentive à analyser une nouvelle image de la pensée. Celle-ci articule trois thèses majeures :

– une thèse sur la situation originaire de signification ou *proto-signifiance*. Celle-ci présente un minimum de complexité qu'on peut exprimer par la conjonction de trois conditions nécessaires sur les trois axes de la *semiosis* : différence sémiotique, référence sémantique, communicabilité pragmatique. C'est faire reculer l'ancienne conviction d'un sens subjectif qui dérive de l'activité d'un *ego* conscient, au profit d'un processus dynamique complexe, que la faute philosophique serait de dissocier.

– une thèse sur le *sens textuel* qui est produit par une élaboration sur chacun des axes de la signifiance. Interrogative dans son ressort et son mouvement, signifiante par ses conditions de possibilité symbolique, la pensée est textuelle dans son élaboration effective et sa portée innovante. Son unité renvoie au questionnement qui donne au texte son mouvement et la capacité de se projeter *ad extra* et d'associer un auteur et un lecteur en les faisant entrer en interrogation. L'existence d'un surcodage par le genre (un texte vaut comme poème ou roman, hymne, compte-rendu d'expérience etc.). Et son organisation en *corpus* plus ou moins homogènes nous conduit à passer de la considération du discours à celle du texte.

– Une conséquence sur la *cognitio*. Si, en effet, la pensée est foncièrement interrogative et soumise à la condition de textualité[17], il faut sans doute revoir nos postulats sur la cognition. Que la fonction calculante de la pensée soit conduite séquentiellement sous la direction d'un centre de contrôle, comme dans le modèle cognitiviste, ou qu'elle soit menée en parallèle par effet d'interaction locale comme dans un réseau connexionniste, de toute façon elle est transférable de l'homme à des ordinateurs rapides et puissants. Le transfert de la pensée calculante peut assurément s'étendre à la représentation des connaissances déjà élaborées. Les ordinateurs manipulent des états de mémoire physique, associés à une représentation symbolique des faits, régie par des codes variés. Mais il manque aux significations purement fonctionnelles qui sont dégagées plusieurs caractéristiques de l'usage de langue naturelle qui subsiste dans la recherche scientifique qui élabore la connaissance : notamment l'auto-référence et l'interrogativité. Celles-ci limitent le transfert des fonctions humaines vers les artefacts de la machine. Ici peut-être se laisse lire, pour le moment, une différence importante entre l'intelligence naturelle et l'intelligence artificielle.

17 On désigne ainsi la condition qui fait dépendre la pensée du processus de signification textuelle, tant pour son instauration que pour sa viabilité et sa spécification.

Il est indiscutable que les technologies du signe, l'informatique appliquée au texte révèlent des structures du signifiant qui échappent au lecteur naturel. Leur analyse est facilitée par la consultation des occurrences de termes. Elle apprend au lecteur une foule de données qu'il ne pouvait maîtriser. Reste à savoir si ces structures sont pertinentes pour l'identité du sens textuel et l'unité de la pensée en exercice. La rationalité procédurale (calculer ce qu'on cherche) n'épuise pas la rationalité. Avant la cognition, il y a le penser, disons la cogitation. Penser, ce n'est pas calculer ni même juger, c'est interroger. Après le *linguistic turn*, l'épistémologie s'est d'abord placée sous le contrôle du langage de la théorie achevée. Depuis une quinzaine d'années, les SC ont stipulé que nos expressions symboliques constituent le dépôt sédimenté de nos opérations cognitives, dont elles sont contemporaines.

Tout dépend de quelle SC on veut parler et à quelle époque épistémologique on se situe. Si c'est d'intelligence artificielle par exemple, il convient de s'interroger sur la notion de *simulation*. Le praticien de l'intelligence artificielle ne vise pas à *créer* un comportement intelligent mais seulement à le *simuler*[18]. Une entité simulante doit certes ressembler à son modèle mais il est contre-productif de pousser la ressemblance trop loin. Les mérites des réalisations de l'intelligence artificielle doivent s'évaluer sur la base des performances. C'est le résultat qui compte. Le fait que le programme simule plus ou moins fidèlement les moyens intellectuels est non pertinent. Il est plus efficace de simuler un comportement intelligent quand on peut évaluer sa qualité avec une certaine précision. Le plus souvent, on cherche à simuler des comportements particuliers dans des contextes restreints.

De plus, je me demande si cette hypothèse, à l'époque cognitiviste en tout cas, ne repose pas sur la seule étude des *contextes de justification*. L'investigation plus récente des *contextes de découverte* tend à mettre l'épistémologie post-popérienne sous le contrôle du langage interrogatif de la recherche plutôt que le langage assertorique de la théorie, comme c'était le cas dans la conception standard. Cela revient, pour penser la cognition, à sortir de la considération de la connaissance toute faite pour rendre leur irréductibilité aux notions de problème et de problématique dans la connaissance.

Ce qui rend une connaissance intéressante est autre chose que le rapport de la connaissance toute faite à son langage : par exemple le rapport logique qu'elle entretient avec une situation de problème, sa dynamique et sa relation aux théories rivales, son aptitude à résoudre

---

18 18 J.R.Searle, 'Minds, brains and programs' in Hofstadter D. et Dennett D.C., The mind's eye, Bantam Books, 1981, s'interroge sur la portée philosophique des efforts de simulation des capacités cognitives sur ordinateur. Selon l'intelli- gence artificielle faible, la valeur de l'ordinateur est celle d'un outil puissant pour formuler et tester les hypothèses. Selon l'intelligence artificielle forte, l'ordinateur convenablement programmé est véritablement un esprit capable de comprendre.

des problèmes existants et à en suggérer de nouveaux. La signification d'une connaissance dépend d'un contexte très étendu, pris entre les problèmes que la connaissance résout et les problèmes qu'elle soulève. Ce qui est intéressant dans la cognition comme processus, c'est le questionnement qui aboutit à élaborer une problématique dans son rapport avec une situation de problème, conformément au mouvement de la pensée interrogative. Nous voici revenus à examiner la modalité de la pensée interrogative à l'œuvre dans tel ou tel type de texte. Les sciences de la *cognitio* – dont fait partie l'épistémologie – rencontrent de manière instructive pour elles les sciences du texte.

3. Philosophie de l'esprit et Science de l'esprit

Le présent débat est une invitation à fournir à l'anthropologie philosophique l'introduction dont elle a besoin. Les SC comme sciences naturelles de l'esprit voudraient abolir la vieille division entre science de l'esprit et de la nature. Les représentations, la conscience, le langage, la catégorisation, la mémoire seraient compris comme des phénomènes naturels, expérimentables et modélisables. Les SC renforcent le courant *naturaliste* même si la naturalisation de l'esprit reste problématique.

Revenons sur quelques aspects du projet 'en voie de réalisation' d'une *science de l'esprit*, pour le confronter à la *philosophie de l'esprit*. Un des problèmes vient de ce que chacune exerce une sorte de droit d'inclusion ou même d'intrusion à l'égard de l'autre. Quelle est, en effet, la place de la connaissance dans l'ordre naturel, de l'acte de connaître dans l'ordre du connu ? Si le penser déborde le connaître, pourquoi n'y aurait-il pas, au-delà des sciences de la cognition au sens le plus technique du mot, des sciences de la *cogitatio* au sens le plus classique ? Si la cognition n'éteint pas la *cogitatio,* l'accent pourrait être mis à égalité sur l'architectonique des sciences bien sûr, mais aussi sur les arts et la poétique, sur les théologies et les philosophies, bref sur toute manifestation de la pensée. Alors pourquoi pas des sciences de la *cogitatio* ? Après cette interrogation qui est plus qu'une boutade, je me demande si la philosophie de l'esprit au sens de la *philosophy of mind* n'était qu'une partie de la philosophie, aurait-elle vocation à entrer dans le polygone de l'interdiscipline des SC.

Question plus ponctuelle : sur l'anti-psychologisme. La psychologie cognitive est-elle hors de ses prises ? Je voudrais rappeler un trait de sa correspondance avec Meinong (1899-1907), qui est un trait de l'histoire de la lutte contre le psychologisme. Entre psychologie et logique, c'est une querelle de priorité, explique Russell : la logique est première parce qu'il faut savoir ce qu'est une définition et un raisonnement pour mettre

en place l'ordre des matières et l'ordre des raisons de la psychologie. Cet argument pour le compte de la logique mathématique vaut d'être rappelé. Il en rejoint un autre pour le compte de la philosophie transcendantale trop souvent assimilée à une psychologie transcendantale. Et si penser n'était pas un concept d'expérience ? Comme dit Wittgenstein... Ce n'est pas parce que nous pouvons nous observer en train de penser que nous savons ce qu'est la pensée.

Questions de seconde espèce : c'est ici la philosophie des sciences qui est concernée, avant la philosophie de l'esprit. La possibilité théorique de *rabattre sur le binôme syntaxe/sémantique le binôme computo-représentationnel* repose sur une condition d'isomorphisme qui n'est réalisée en toute rigueur que pour les systèmes formels les plus simples. On sait que le modèle de Turing étend la notion de mécanisme au paradigme de la *computatio*. Mais raisonner n'est pas calculer (réductible à un calcul algorithmique). La syntaxe ne peut jusqu'au bout simuler la sémantique. D'autant que les propriétés sémantiques d'un état cognitif dépendent des relations avec l'environnement externe. L'analogie technico-théorique de l'ordinateur ne peut rendre compte jusqu'au bout de l'efficacité causale des états mentaux et de leur sémanticité ? Une fois 'transposée' du fonctionnalisme computationnel à la sémantique des logiciens la situation d'isomorphisme que requiert un modèle à la *Turing* peut donc faire difficulté. Elle ne vaut pas au même sens pour le calcul des prédicats monadiques et pour les systèmes d'ordre plus élevé qui ne sont plus 'décidables'.

Je réagis ensuite sur la chronométrie des temps de réaction. Comment retrouver l'homogénéité des unités de mesure ? A supposer que je puisse mesurer le temps mis pour comprendre le deuxième axiome de Peano et le temps mis pour comprendre le second mouvement d'une symphonie de Mozart, y a-t-il un sens à dire que leur chronométrie est à peu près semblable pour *qualia* aussi divers ? Je me sens brusquement kantien, en demandant à la psychologie de définir une métrique qui engage son rapport à l'égard de l'espace et du temps. Les schémas connexionnistes soulèvent une perplexité analogue.

Wittgenstein estimait que la psychologie fait cohabiter des méthodes expérimentales et une confusion dans les concepts. Quand on pense que le terme *représentation* évoque les problèmes d'intentionnalité, les problèmes épineux de rapport à l'environnement externe, je me dis que, décidément, c'est un terme un peu léger pour un terme philosophiquement aussi lourd.

Dernière question sur le modèle théorique qui engage l'ontologie des entités retenues : il me semble que l'argument de *supervenience* affronte la problématique de l'existence des *normes*. C'est vrai, rien n'existe dont la science ne finisse par attester l'existence : particule, électron, cellule, pourquoi pas norme ? Mais si, comme disait Bachelard, l'existence est

une fonction non monotone, alors c'est toute philosophie qui revient, convoquée pour réfléchir : est-ce de la même existence qu'on parle pour une norme, une cellule, une relation ou un être mathématique ?

Il importe aussi de savoir dans quel sens on va parler des fondements naturels de l'éthique, comment on va poser la question de l'héritage biologique et culturel des normes morales[19]. L'équivoque se concentre sur la notion même de *fondement*. Ou on vise les *bases* d'un édifice, comme sous-sol sur lequel on veut pouvoir construire, mais sans que la compréhension du sous-sol prétende nous offrir la compréhension du bâtiment. Il y a d'autre part le fondement au sens du *principe*, dans la philosophie de l'expérience morale, où le biologique représente un niveau partiel de sous-bassement, tel que le surgissement d'un nouveau niveau de sens soit intelligible à partir de lui.

La pensée, le sens, la représentation etc. ne sont justement pas des 'problèmes'. Ce ne sont pas non plus des 'arguties sans intérêt' mais des questions radicales, ouvertes, informelles dont on ne veut pas hypothéquer la forme de réponse possible, à l'intérieur d'un modèle théorique. Que la philosophie donne peu de réponses ne vient pas d'une complaisance pour la perplexité mais simplement de la nature de ses questions. On peut soutenir, me semble-t-il, que la pensée du philosophe en son érotétique radicale -- comme la pensée du poète en son énigmatique, comme *l'intellectus fidei* du théologien quand il tente d'élucider le mystère -- diffère de la pensée des sciences positives en terme de problème. Elle n'existe que dans la mesure où nous tenons compte du fait que le sens ultime de notre rapport à l'inconnu nous échappe. Je doute qu'une question philosophique soit reformulable sans reste en termes de problème scientifique. Pour une raison qui tient à la logique des questions et des réponses : les questions 'suscitées en philosophie de l'esprit' le seront *modulo* une reprise catégoriale qui risque d'oblitérer le mode d'interrogativité en cause. Il en va de l'esprit, disons de la *cogitatio* dans la multiplicité des modes de sa compétence interrogative.

On a vu la vigilance épistémologique du phénoménologue s'exercer à l'égard des homologations disciplinaires. Il m'apparaît maintenant que la tendance hégémonique des disciplines est somme toute moins contestable que la tendance hégémonique qui pousse à oblitérer les autres modes d'interrogation ou compétences de pensée qui président à la recherche sur les grandes méta-questions, en l'occurrence sur la représentation, mais aussi sur le langage, la compréhension, l'informa-

---

19 Le biologiste naturaliste n'est pas obligé de réduire simplement chaque niveau de l'évolution au niveau inférieur, mais il interprète les effets de seuil dans une perspective qui reste strictement matérialiste. Il utilise la théorie des systèmes pour expliquer que 'des qualités qualitativement nouvelles apparaissent à chaque niveau d'organisation du simple fait que les éléments constitutifs peuvent coopérer entre eux'(p. 269).

tion, toutes questions fortement amphibologiques. Il est encore moins possible de transformer un problème de frontières *intercompétentiel* en problème *intracompétentiel*, que de transformer un problème *interdisciplinaire* en problème *intradisciplinaire*. A l'horizon, l'enjeu ultime : ce que nous objectivons de l'esprit est-il l'esprit ?

La réflexion contemporaine en philosophie de l'esprit et surtout en philosophie des sciences est liée au développement des SC. Il n'y a pas seulement conflit de compétences, il y a intrication et convergence. C'est la nature du 'lien' qui est le dernier enjeu de la controverse. Or, je ne vois pas qui sera compétent en dernier ressort pour en décider. Bien entendu, les remarques précédentes ne disposent pas de l'ordre des questions, elles ne font que jalonner le véritable questionnement qui resterait à mettre en place dans une controverse méta-compétentielle en bonne et due forme.

Francis JACQUES

LIFE AND MIND

MARGARET A. BODEN
University of Sussex, U.K.

Abstract: It's sometimes said, and even more often assumed, that life is necessary for mind. If so, and if A-Life promises to throw light on the nature of life as such, then A-Life is in principle highly relevant to the philosophy of mind and cognitive science. However, very few philosophers have attempted to argue for the relation between life and mind. It's usually taken for granted. Even those (mostly in the Continental tradition, including some with a following in A-Life) who have insisted on the linkage have stated it rather than justified it. If an evolutionary account of intentionality is acceptable, then perhaps biological life 'makes room' for mind. But that claim is problematic, since it's not clear that the type of self-organization involved in life-as-such must necessarily include evolution. Even if it does, it's a further step to show that life is strictly necessary for mind.

Key words: Life, self-organization, evolution, intentionality, A-Life. Margaret A. Boden Research Professor of Cognitive Science University of Sussex

I: Introduction

When Alan Turing (1950) claimed, in the august pages of Mind, that there's no good reason to deny that computers might be able to think, philosophers were quick to disagree.

Most of them focussed on the concept of thought. Wolfe Mays (1952), for instance, penned the argument that John Searle (1980) would later express as "all syntax and no semantics", adding for good measure that thought implies consciousness (an addition that led Gilbert Ryle to refuse

to publish Mays' piece in Mind: W. Mays p.c.). But a few appealed also to the concept of life. That is, they relied on the intuition that life is necessary for mind.

Michael Scriven, for example, supplemented the claim that thought implies consciousness, by declaring that "Life is itself a necessary condition of consciousness", and that "Robots ... are composed only of mechanical and electrical parts, and cannot be alive" (Scriven 1953: 233). However, he didn't say why something made of mechanical and electrical parts can't be alive, whereas something made of biochemical parts can (but see the discussion of metabolism in Section III).

Nearly twenty years later, by which time the notion that computers might be capable of thought was no longer surprising (though many people found it no less shocking), Peter Geach said something very similar. AI systems, he insisted, can't have beliefs and intentions because they're "certainly not alive" (Geach 1980: 81).

However, neither Geach nor Scriven tried to explain why life is necessary for thought, or for consciousness. And only Scriven offered a reason why robots "cannot" be alive (namely, that they're made of mechanical and electrical parts). As for Geach, his confident "certainly not alive" wasn't glossed in any way: he deemed it obvious not only that mind depends on life but that computers aren't living things. Their reticence wasn't unusual. The same two claims are made fairly often, but with scant argumentation attached.

An apparent exception, if one gets no further than reading the titles in a bibliography, is Hilary Putnam's paper on "Robots: Machines or Artificially Created Life?" (1964). Indeed, the question of the possibility of living robots was here 'twinned' with that of the possibility of robot minds, since Putnam was responding to Paul Ziff's (1959) essay on "The Feelings of Robots".

However, Putnam's paper focussed mainly on consciousness, not life. At one point, he endorsed Ziff's claim that it's an "undoubted fact" that if a robot isn't alive then it can't be conscious. But he ascribed this truth to "the semantical rules of our language", not to any quasi-explanatory relationship between life and mind. He also said (this time, disagreeing with Ziff) that something which is clearly a mechanism might be alive. Again, however, this was linguistic philosophy in action. Elsewhere, Putnam (1962) had heretically recommended changes in word-meaning due to new scientific data (about dreaming, for instance). But in the paper on robots and life, he relied on what current usage allowed one to say without contradiction. The nearest he got to discussing a substantive claim about life was to scorn the suggestion that the primary difference between a robot and a living organism is the "softness" or "hardness" of the body parts (1964: 691).

It's hardly surprising that Putnam's desultory discussion didn't lead his fellow philosophers to take an interest in the issue of life-and-mind. Most

of them simply took it for granted that these two concepts, or phenomena, are somehow linked--which would imply that if computers aren't alive then they can't be psychological systems. (That's why, in their unquestioned assertions about the relations between life, mind, and computers, Scriven and Geach--and Ziff, too--evidently expected immediate assent.)

The same is true today. Analytic philosophers appear to think the life-mind linkage so obvious that, even when they bother to state it explicitly, they don't offer any arguments for it. Those few philosophers who have provided arguments are on, or near to, the Continental side of the intellectual fence. But even Continentals often state the linkage rather than challenging and justifying it (see Section II).

My purpose here isn't to overthrow the Geach-Scriven intuitions. For I, too, see computers/robots as quintessentially non-living things. And I, too, suspect--though with less confidence--that mind requires life. Rather, I want to examine the reasons for both of these commonly-held beliefs.

Quite apart from the desirability of philosophical hygiene ("No unexamined assumptions, please!"), this relates to the importance of A-Life, as contrasted with AI, for the philosophy of mind.

The oft-declared 'opposition' between AI and A-Life is in fact largely spurious (Boden 2006: chs. 4.ix and 15). But there is a clear distinction of research focus. Whereas AI is the computer-based study of psychology (especially human psychology), A-Life concentrates on ethology and biology--including the nature of life as such. It follows that if the Geach-Scriven intuitions are both well-founded, then A-Life cannot generate mind but is, in principle, relevant for understanding it.

II: The life-mind linkage defended

Proponents of the life-mind linkage include the existentialist theologian Hans Jonas, and the neurophysiologists Humberto Maturana and Francisco Varela (see Boden 2006: 15.viii.b and 16.x.a and c). Unlike Scriven, Geach, Ziff, and even Putnam, they wrote about this matter at some length.

Jonas wasn't interested in biology for scientific purposes, but approached it in an ethical-theological spirit. In his view, orthodox (molecular and experimental) biology should be replaced by a biological science of a very different kind, because it illustrates the disastrous cultural influence of Cartesianism.

Descartes, he complained, had separated human beings from the rest of Nature, by means of a "spiritual denudation" of the non-human world which had stripped it of any intrinsic value (Jonas 1966: 58-63, 232). Jonas believed that this had helped lead his ex-teacher Martin Heidegger toward Nazism, through attaching more importance to the fact that

humans can make free decisions than to considering which values should guide our decisions (Jonas 1990: 200). Those values, he said, are themselves grounded in Nature. He meant not only that our values emerge as a result of Nature, but also that Nature is in itself valuable--and therefore worthy of respect (he later became a guru of the environmentalist movement: see Jonas 1984.)

More specifically, values are grounded in life. Embodiment, and in particular metabolism, was seen by Jonas as philosophically crucial (1966: 64-91). So was evolution. Charles Darwin, he said, despite his materialist assumptions, had helped us to understand that all forms of mediation between organism and environment--perception, motor action, emotion, conscious imagination, and self-reflection--emerge as a result of evolution (1966: 38-58). In general, life is essential for the emergence of mind (99-107).

Indeed, mind is present, or rather prefigured, in all of life. According to Jonas, all self-organized matter is, in a sense, ensouled: life involves "self-concern".

He expressed the life-mind linkage thus: "One way of interpreting [the ascending scale of life] is in terms of scope and distinctness of experience, of rising degrees of world perception.... Another way, concurrent with the grades of perception, is in terms of progressive freedom of action.... [The] 'mirroring' of the world becomes ever more distinct and self-rewarding, beginning with the most obscure sensation somewhere on the lowest rungs of animality, even with the most elementary stimulation of organic irritability as such, in which somehow already otherness, world, and object are germinally 'experienced', that is, made subjective, and responded to" (1966: 2).

Whereas freedom, for Descartes, was a God-like immaterial power, for Jonas it is founded in our biology--indeed, in "the blind automatism of the chemistry carried on in the depths of our bodies" (1966: 5). But that chemistry, he said, differs from the chemistry of "suns, planets, and atoms" in being embodied as metabolism. The "principle of freedom" common to all living organisms lies in their having a special type of identity and continuity: a stable dynamic form made of an ever-changing material substrate. This both enables and prefigures the human capacity for making decisions: "One expects to encounter [talk of freedom] in the area of mind and will, and not before: but if mind is prefigured in the organic from the beginning, then freedom is. And indeed our contention is that even metabolism, the basic level of all organic existence, exhibits it: that it is itself the first form of freedom" (1966: 3).

In short, "mind is prefigured in organic existence as such" (5). Life and mind are ontologically inseparable: "the organic even in its lowest forms prefigures mind, and ... mind even on its highest reaches remains part of the organic" (1).

It's evident from these quotations that, besides trying to explain why all the minds we know about are found in living things, Jonas was trying also to say what life is.

The same applies to Maturana and Varela (1980), who defined life as "autopoiesis in the physical space". This concept is close to, though not identical with, metabolism (Boden 2000). Autopoiesis in general, they said, is the continuous self-production of an autonomous entity. The boundaries, components, and internal relations of "an autopoietic machine" (i.e. a living organism) are produced and maintained by a network of self-organizing processes (1980: 79). The system thus "pulls itself up by its own bootstraps and becomes distinct from its environment through its own dynamics, in such a way that both things are inseparable" (ibid.).

As biologists, Maturana and Varela had a much better grasp than Jonas did of the scientific issues involved (although their prose was even more convoluted and obscure than his). They were less concerned with ethics, and more with the adaptive functionality of organisms. (This isn't to say that they were functionalists: in theorizing the mind/brain, they explicitly rejected talk of input, output, computation, internal representations, and even feature-detectors--which had been discovered by a team including Maturana himself: Lettvin et al. 1959.) Accordingly, where Jonas had credited all living things with "freedom" and "self-concern", they credited them instead with "cognition"--calling their major book "Autopoiesis and Cognition".

All three terms, in my view, should be taken with a large pinch of salt when associated with the concept of life. There is indeed an important sense in which living things, at base thanks to metabolism, have a degree of autonomy. But to call this "freedom", or even a "prefiguration" of human free choice, demands a far more detailed justification than Jonas provided.

Similarly, living things are, indeed, pre-adapted to the specifics of their habitat, and capable (in varying degrees) of adapting to it appropriately when it changes. To call this "cognition", however, is to push the term too far. When an oak-tree loses its leaves in the autumn, this is an adaptive response, nicely adjusted to the environmental conditions. But knowledge (cognition), properly so called, it is not. What ethologists term innate releasing mechanisms, such as the hawk-like stimulus that prompts the newly-hatched grouse chick to crouch, are somewhat more persuasive. Even so, the longstanding philosophical debate about the coherence of the notion of "innate knowledge" (e.g. Edgley 1970) shows how problematic it is to ascribe knowledge to creatures (newborn babies, as well as grouse) lacking evidence, hypothesis, judgment, or even learning.

However, to say that a claim should be taken with a pinch of salt is not to throw it off the dining-table altogether. For one thing, linguistic usage can be extended in the light of scientific advance--as Putnam (1962) had

rightly insisted. So terms such as "freedom" or "knowledge" might justifiably come to be ascribed way beyond their usual boundaries. (And recent empirical research, while showing that the seemingly innocuous term "innate" is in fact highly problematic, has provided ample data for various types of prefigurement in newborn animals: Elman et al. 1996.)

More to the point, for present purposes, an unconvincing argument is better than no argument at all. The three writers I've mentioned here did at least make a start in explaining the link between life and mind, instead of taking it for granted.

III: What is Life?

As we've seen, part of their project (besides establishing the life-mind link) was to say what life is. There's still no universally agreed definition of it. Indeed, one of the aims of A-Life research, as expressed by Chris Langton (1989), is to arrive at one. Nevertheless, ten characteristics are mentioned repeatedly in attempts at defining life: self-organization, autonomy, emergence, development, adaptation, responsiveness, evolution, reproduction, growth, and metabolism.

One could say that only one characteristic is crucial, for self-organization (the spontaneous appearance of new levels of order) covers all the others as special cases. It's no accident, then, that Jonas and Maturana and Varela referred continually to self-organization in their discussions of life.

One could also say--and A-Life researchers typically do so--that the first eight characteristics listed above are abstract (functionalist) notions, defining aspects of what's sometimes called the logical form of life. As such, they are grist to the mill of a computer-based approach. Even growth might be included, if we allow that this term is ambiguous as between physical growth and incresase in size-otherwise-defined (e.g. as length/number of program instructions).

Only metabolism--also heavily stressed by the writers featured in Section II--remains incalcitrant to an abstract, functionalist, interpretation. For it's irredeemably physical. Moreover, it doesn't refer only to energy-use, or even to individual energy-budgeting (both of which can be ascribed literally to computers). Rather, it refers to the use and budgeting of energy in the autonomous construction and maintenance of the living system itself. This, given evolution, will inevitably involve a set of interacting biochemical cycles, of increasing complexity (Boden 1999).

That's why strong A-Life, i.e. virtual life, is impossible. For computers don't metabolize, in the sense just defined. (It doesn't follow that A-Life can't throw significant light on various examples of self-organization,

metabolism included. Indeed, it has already done so--see Boden 2006: ch. 15.)

However, these remarks don't settle our query here. For while self-organization is undeniably a key concept in defining life, it's not at all clear that it's a key feature of mind. This is a conceptual point, not an empirical one; so the growing evidence of self-organization in the brain, though fascinating, is irrelevant (Boden 2006: 14.vi.b and ix.a-d). Only if we were to define mind as mind/brain would this evidence be conceptually germane.

Nor is it clear, pace Jonas, that mind must arise from metabolism. One might say that mind, and intelligence, is necessarily adaptive. But does this mean anything over and above 'efficient', and/or 'well-suited to the specifics of the environment', and/or 'capable of change through learning'? All of those descriptions could conceivably be attached to robots.

In particular (and, again, pace Jonas), it's not obvious that minds must necessarily be generated by evolution. It's not even obvious that life itself must involve evolution.

Maturana and Varela, for instance, argue that evolution (and reproduction too) presupposes autopoiesis, so can't be essential to it (1980: 105ff.). Moreover, to include evolution in the list of vital characteristics has some counterintuitive implications (Bedau 1996). Since the concept applies only to populations, an individual oak-tree or lion--usually regarded as paradigm cases of life--can't be regarded as a living thing except by appealing to their ancestry. A non-evolving population, temporarily in evolutionary equilibrium, wouldn't exemplify life either. And creationism becomes incoherent, not simply false.

Nevertheless, all the living things we know of have evolved. Moreover, it's difficult (to put it mildly) to see how highly complex organisms could be generated except via evolution--a point which even Maturana and Varela were happy to grant. In addition, the concept of evolution enables us to inter-relate all known living things, and to explain a host of details about them--which is why it's so often included in the definition of life.

Let's agree, then, that all life (with the possible exception of the most primitive autopoietic unities) has in fact evolved, whether or not we also choose to say that it must have done so. Does this have any bearing on the life-mind linkage?

IV: Evolution and intentionality

Talk of evolution reminds us of another way of linking life and mind. For intentionality, or meaning, is the key aspect of mind, and some philosophers, such as Ruth Millikan (1984) and David Papineau (1987), have argued that intentionality is grounded in evolution. (So did Jonas, as

we've seen; but whereas he asserted it in a rhetorically persuasive fashion, they considered counter-arguments carefully in stating their case.) In a nutshell, what these authors argue is that it is adaptive function, favoured and fixed by evolution, which gives meaning to animals' actions--including the linguistic activities of human beings.

I said, in Section III, that it's not obvious that minds must be generated by evolution. Not only isn't it obvious (i.e. clear at first sight), but it's still widely contested. Wittgensteinian philosophers, for example, deny the possibility of any naturalistic account of meaning, evolutionary or not (e.g. Morris 1992; McDowell 1994). (Hence Millikan's cheeky book-title, outrageously describing language and thought as "biological" categories.) Moreover, even when the claim is accepted, as the best alternative on offer, it's usually admitted to be problematic.

For instance, even sympathetic readers of Millikan's book may be discomfited by her "swamp-man" thought-experiment (Millikan 1984: 93, 337f.; 1996). She admits that her philosophy of intentionality implies that a magically constituted molecule-for-molecule copy of Jo Blogg's body and brain would have no knowledge, no beliefs, no understanding ... despite responding to all our greetings and questions exactly as Bloggs would have done. In short: no evolution, no meaning. She's discomfited by this implication herself. But she refuses to be defeated by it, arguing that the thought experiment is so utterly unrealistic that it's not worth taking seriously. I agree with her. (Similarly, I wouldn't drop thermodynamics simply because it allows for the theoretical possibility, if only for a split second, of a snowball in Hell.) But not all philosophers would.

If the anti-naturalists are right, and even biology (evolutionary theory and/or neurophysiology) can't explain intentionality, then computers certainly can't do so. And if they're wrong, it still doesn't follow that a computer-based discipline can help supply the naturalistic explanation being sought. We've already seen that computers aren't alive. Can programs, or robots, nevertheless be said to evolve? And if so, can they help us to understand the origin of meaning?

Evolution is the gradual change of a population whose individual members reproduce ('asexually' or 'sexually') with inheritance and variation, where some fitness function selects the next breeding-individuals at each generation. The "change" is typically an improvement, with respect to the 'task' implied by the fitness function. Given the abstract nature of this concept (see Section III), the systems concerned need not be organisms: they may be programs, or even robots.

Programs evolve by employing genetic algorithms, or GAs. These enable reproduction--either self-copying or 'copying' from two parents--with variation. They make random changes (mutations, crossovers ...) in some of the program-rules at the copying stage, and then apply a fitness function to select the 'best' resulting individuals (plus a few others, to

allow potentially advantageous variations to remain in the 'gene pool'). (Sometimes, it's the programmer who applies the fitness function; for our purposes, however, the fully automatic cases are more relevant.)

GAs can be used to tackle non-biological tasks, such as optimizing the design of an aircraft engine. But A-Life researchers are typically concerned with biological phenomena. Not all A-Life research involves GAs: much of it studies non-evolutionary typoes of self-organization. But much of it does, and the examples mentioned below are all taken from 'evolutionary' A-Life.

Usually, the researchers pick a single 'task' and try to evolve programs that can achieve it, in one or more environments. But, as in biology itself, "evolution" may involve co-evolution of distinct groups, or species--predator and prey, for instance (Cliff and Miller 1995). The first, and still the best-known, example is Tom Ray's (1992) Tierra model. This program is famous partly because of its results: to Ray's amazement, it generated parasites, counter-parasites (i.e. hosts resistant to the previous parasites), and super-parasites (which overcame the hosts' previously evolved resistance); it also proved that gradual genetic change can underlie what looks like saltatory, or "punctuated", evolution. It's famous also because of Ray's claim that virtual "creatures" like those evolved by Tierra are genuinely alive. (He counters the 'no metabolism' objection by pointing out that computers use energy; however, we've seen that metabolism involves much more than this.)

As for robots, the central controller, or neural-network 'brain', can be evolved from a starting-point wherein the connections between the component units (and their nature: excitatory or inhibitory) are random. Similarly, sense-organs--such as whiskers, or the second eye--that aren't actually needed for the task may lose their connection to the 'brain' and become useless, rather like the human appendix (Cliff, Harvey, and Husbands 1993). Or their anatomy may evolve, so that predators develop narrow-angle, forward-looking, eyes, while the prey develop wider-range, front-and-sideways, vision: think foxes and rabbits, respectively (Cliff and Miller 1995).

What has this to do with intentionality? Two things. On the one hand, if meaning is grounded in evolution then life--which, we've agreed, involves evolution (whether necessarily or as a matter of fact)--is at least 'suitable' for mind. To that extent, the life-mind linkage is supported. On the other hand, non-living but genuinely evolved robots may offer a foothold for the ascription of meaning that isn't available to other computer systems.

Searle directed his "all syntax and no semantics" objection to robots, as well as to programs. He said--and he was right--that the 'meanings' we ascribe to program-instructions come wholly from us. In principle, one and the same program could be interpreted either in terms of the tax-laws or in terms of dance-steps. There is nothing intrinsic to the program to choose between the two interpretations (or three, or ...). With robots, it's a

bit more tricky. After all, it might be a dancing robot, controlled by the tax-or-choreography program--in which case the second interpretation would seem to be clearly more appropriate. Even here, however, Searle could point out that the fact that the robot performed such-and-such a step at a certain time is wholly dependent on its programmer: he/she programmed it to dance the polka, but could have programmed it to tap out the numbers for your tax-return.

With an evolved robot, however, the case is different. Let's take an actual example. A population of robots was being evolved to navigate a particular environment so as to reach a certain spot, irrespective of the starting-point (Harvey, Husbands and Cliff 1994; Husbands, Harvey and Cliff 1995). One of the things in the environment was a triangle of white cardboard. Sometimes (for evolution is probabilistic, not deterministic), the neural-network controller evolved a 'feature detector' analogous to those discovered in monkeys' brains. This was a mini-network sensitive to a light-dark gradient at a particular orientation. (No other such mini-networks evolved in this case; so a black triangle, or the right side of the white triangle, were in effect invisible.)

Putting this example in intentional terms, activity in the mini-network meant 'light-dark gradient at orientation x', or perhaps 'left side of white triangle', or even 'landmark directing me to veer to the right'. Since the robot lacked language, and hadn't even evolved a rich set of visual discriminations, it's not obvious which of these descriptions one should pick. In other words, if there was content (meaning) here, it was non-conceptual content (which some philosophers deny outright, even in non-human animals: McDowell 1994).

The point, however, is that the mini-network evolved as part of a visuomotor mechanism. The connections of specific visual units in the 'brain' to specific motor units enabled the robot to use the white triangle as a navigation aid in achieving the task which (unknowingly) it had been set. Indeed, the feature-detector's very existence, as well as its function, depended on that evolutionary history--not on any foresight by the roboticists.

Even if one doesn't accept an evolutionary account of meaning in general, it's surely more appropriate to ascribe one of the three meanings suggested above to the mini-network than to say--inspired by Searle--that it could mean just anything. Certainly, it couldn't be used as a tax-calculator: that would require its components and connections--its nature--to be very different. And if one does see evolutionary history as essential to meaning, this conclusion is even firmer.

My claim, here, isn't that an evolutionary semantics must ascribe real (non-conceptual) meanings to this robot, or to complexified versions of it. For although the existence of that specific mini-network depends wholly on evolution, the existence of the (initially, random) neural-network controller itself does not. On the contrary, it depends on deliberate human

agency. One might wish to argue that this makes all the difference, that "real" meaning demands evolution all the way down. I shan't pursue that point here (but see Boden 1972: 119, 195). It's enough, for my purposes, to show that if one does ascribe meaning (whether real or metaphorical) to the robot and/or to its feature-detector then there are strict limits on the meanings one can plausibly suggest. Orientation-detector, yes; navigational landmark, probably; one side of a white triangle, perhaps; tax-calculations, certainly not.

V: Conclusion

I've argued that life implies evolution (though perhaps not of necessity). And I've suggested that evolution-based philosophies of meaning are the most plausible accounts on offer (and that we need some naturalistic semantics or other). If both those claims are true, then life is well-suited for mind, even if it's not actually necessary for it.

Computers, robots included, aren't living things, because they don't metabolize in the required sense. But some of them do, genuinely, evolve. Of those, some are systems to which we naturally ascribe meanings which are not arbitrary with respect to the artefact itself. On the contrary, they are grounded in its specific evolutionary history.

Whether these (non-arbitrary) "meanings" are real or merely metaphorical depends partly on whether our philosophical semantics demands evolution all the way down. For even a robot of the ten-millionth generation, whose behaviour (and anatomy) was unforeseen, wouldn't have existed if human beings hadn't embarked on evolutionary robotics in the first place.

Finally, this sub-area of A-Life may help us to model the evolution of non-conceptual content, and (with language added) possibly of conceptual content too. In so doing, and even if that content is merely metaphorical, it may help to clarify our philosophical ideas about mind, and its relation to life.

References:

Bedau, M. A. (1996), 'The Nature of Life', in M. A. Boden (ed.), The Philosophy of Artificial Life (Oxford: Oxford University Press, 1996), pp. 332-357.

Boden, M. A. (1999), 'Is Metabolism Necessary?', British Journal for the Philosophy of Science, 50: 231-248.

Boden, M. A. (2000), 'Autopoiesis and Life', Cognitive Science Quarterly, 1: 1-29.

Boden, M. A. (2006), Mind as Machine: A History of Cognitive Science (Oxford: Clarendon Press).

Cliff, D., Harvey, I., and Husbands, P. (1993), 'Explorations in Evolutionary Robotics', Adaptive Behavior, 2: 73-110.

Cliff, D., and Miller, G. F. (1995), 'Tracking the Red Queen: Measurements of Adaptive Progress in Co-evolutionary Simulations', in F. Moran, A. Moreno, J. J. Merelo, and P. Chacon (eds.), Advances in Artificial Life: Proceedings of the Third European Conference on Artificial Life (ECAL95) (Berlin: Springer-Verlag), pp. 200-218.

Edgley, R. (1970), 'Innate Ideas', in G. Vesey (ed.), Knowledge and Necessity (London: Macmillan), pp. 1-33.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996), Rethinking Innateness: A Connectionist Perspective on Development (Cambridge, Mass.: MIT Press).

Geach, P. T. (1980), 'Some Remarks on Representations', Behavioral and Brain Sciences, 3: 80-81.

Harvey, I., Husbands, P., and Cliff, D. (1994), 'Seeing the Light: Artificial Evolution, Real Vision', in D. Cliff, P. Husbands, J.-A. Meyer and S. W. Wilson (eds.), From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior (Cambridge, Mass.: MIT Press), pp. 392-401.

Husbands, P., Harvey, I. and Cliff, D. (1995), 'Circle in the Round: State Space Attractors for Evolved Sighted Robots', Journal of Robotics and Autonomous Systems, 15: 83-106.

Jonas, H. (1966), The Phenomenon of Life: Toward a Philosophical Biology (New York: Harper Collins). Page-references are to the 2001 edition: Evanston, Illinois: Northwestern University Press.

Jonas, H. (1984), The Imperative of Responsibility: In Search of an Ethics for the Technological Age (Chicago: Chicago University Press).

Jonas, H. (1990), 'Heidegger's Resoluteness and Resolve: An Interview', in G. Neske and E. Kettering (eds.), Martin Heidegger and National Socialism: Questions and Answers (New York: Paragon House), pp. 197-203.

Langton, C. G. (1989), 'Artificial Life', in C. G. Langton (ed.), Artificial Life (Redwood City, CA: Addison-Wesley), pp. 1-47. Revised version in M. A. Boden (ed.), The Philosophy of Artificial Life (Oxford: Oxford University Press, 1996), pp. 39-94.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959), 'What the Frog's Eye Tells the Frog's Brain', Proceedings of the Institute of Radio Engineers, 47:11 (1959), 1940-1959.

McDowell, J. (1994), Mind and World (Cambridge, Mass.: Harvard University Press).

Maturana, H. R., and Varela, F. J. (1980), Autopoiesis and Cognition: The Realization of the Living (Boston: Reidel).

Mays, W. (1952), 'Can Machines Think?', Philosophy, 27: 148-162.

Millikan, R. G. (1984), Language, Thought, and Other Biological Categories: New Foundations for Realism (Cambridge, Mass.: MIT Press).

Millikan, R. G. (1996), 'On Swampkinds', Mind and Language, 11: 105-117.

Morris, M. R. (1992), The Good and the True (Oxford: Clarendon Press).

Papineau, D. (1987), Reality and Representation (Oxford: Basil Blackwell).

Ray, T. S. (1992), 'An Approach to the Synthesis of Life', in C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (eds.), Artificial Life II (Redwood City, CA: Addison-Wesley), pp. 371-408.

Scriven, M. (1953), 'The Mechanical Concept of Mind', Mind, 62: 230-240.

Searle, J. R. (1980), 'Minds, Brains, and Programs', Behavioral and Brain Sciences, 3: 417-457.

Ziff, P. (1959), 'The Feelings of Robots', Analysis, 19: 64-68.

BEYOND MIND.
HOW BRAINS MAKE UP ARTIFICIAL COGNITIVE SYSTEMS

*Lorenzo Magnani, Department of Philosophy and Computational Philosophy Laboratory, University of Pavia, Pavia,Italy, phone +39-0382-986279, and Department of Philosophy, Sun Yat-sen University, Guangzhou (Canton), P.R. China,*lmagnani@unipv.it

Abstract: What I call *semiotic brains* are brains that make up a series of signs and that are engaged in making or manifesting or reacting to a series of signs: through this semiotic activity they are at the same time engaged in "being minds" and so in thinking intelligently. An important effect of this semiotic activity of brains is a continuous process of disembodiment of mind that exhibits a new cognitive perspective on the mechanisms underling the semiotic emergence of meaning processes. Indeed at the roots of sophisticated thinking abilities there is a process of disembodiment of mind that presents a new cognitive perspective on the role of external models, representations, and various semiotic materials. Taking advantage of Turing's comparison between "unorganized" brains and "logical" and "practical" machines" this paper illustrates the centrality to cognition of the disembodiment of mind from the point of view of the interplay between internal and external representations, both mimetic and creative. The last part of the paper describes the concept of *mimetic mind* I have introduced to shed new cognitive and philosophical light on the role of computational modeling and on the decline of the so-called Cartesian computationalism.

Turing Unorganized Machines

Logical, Practical, Unorganized, and Paper Machines

Aiming at building intelligent machines Turing first of all provides an analogy between human brain and computational machines. In "Intelligent Machinery", written in 1948 (Turing, 1969), he maintains that "the potentialities of human intelligence can only be realized if suitable education is provided" (p. 3). The concept of *unorganized machine* is then introduced, and it is maintained that the infant human cortex is of this nature. The argumentation is indeed related to showing how such machines can be educated by means of "rewards and punishments".

Unorganized machines are listed among different kinds of existent machineries:

- *Universal) Logical Computing Machines (LCMs)*. A LCM is a kind of discrete machine Turing introduced in 1937 that has

> [… an infinite memory capacity obtained in the form of an infinite tape marked out into squares on each of which a symbol could be printed. At any moment there is one symbol in the machine; it is called the scanned symbol. The machine can alter the scanned symbol and its behavior is in part described by that symbol, but the symbols on the tape elsewhere do not affect the behavior of the machine. However, the tape can be moved back and forth through the machine, this being one of the elementary operations of the machine. Any symbol on the tape may therefore eventually have innings (Turing, 1969, p. 6)

This machine is called Universal if it is "such that if the standard description of some other LCM is imposed on the otherwise blank tape from outside, and the (universal) machine then set going it will carry out the operations of the particular machine whose description is given" (p. 7). The importance of this machine resorts to the fact that we do not need to have an infinity of different machines doing different jobs. A single one suffices: it is only necessary "to program" the universal machine to do these jobs.

- *(Universal) Practical Computing Machines (PCMs)*. PCMs are machines that put their stored information in a form very different from the tape form. Given the fact that in LCMs the number of steps involved tends to be enormous because of the arrangement of the memory along the tape, in the case of PCMs "by means of a system that is reminiscent of a telephone exchange it is made possible to obtain a piece of information almost immediately by 'dialing' the position of this information in the store" (p. 8). Turing adds that "nearly" all the PCMs under construction have the fundamental properties of the Universal Logical Computing Machines: "given any job which could have be done on an LCM one can also do it on one of these digital computers" (*ibid.*) so we can speak of Universal Practical computing Machines.

- *Unorganized Machines.* Machines that are largely random in their constructions are called "Unorganized Machines": "So far we have been considering machines which are designed for a definite purpose (though the universal machines are in a sense an exception). We might instead consider what happens when we make up a machine in a comparatively unsystematic way from some kind of standard components. […] Machines which are largely random in their construction in this way will be called 'Unorganized Machines'. This does not pretend to be an accurate term. It is conceivable that the same machine might be regarded by one man as organized and by another as unorganized." (p. 9). They are machines made up from a large number of similar units. Each unit is endowed with two input terminals and has an output terminals that can be connected to the input terminals of 0 or more of other units. An example of the so-called unorganized A-type machine with all units connected to a synchronizing unit from which synchronizing pulses are emitted at more or less equal intervals of times is given in Figure 1 (the times when the pulses arrive are called moments and each unit is capable of having two states at each moment). The so-called A-type unorganized machines are considered very interesting because they are the simplest model a

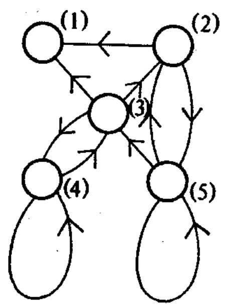
nervous system with a *random arrangement of neurons* (cf. the following section "Brains as unorganized machines").

Figure 1. (In Turing, 1969).

- *Paper Machines*. "It is possible to produce the effect of a computing machine by writing down a set of rules of procedure and asking a man to carry them out. […] A man provided with paper, pencil and rubber, and subject to strict discipline, is in effect a universal machine" (p. 9). Turing calls this kind of machine "Paper Machine".

Continuous, Discrete, and Active Machines

The machines described above are all *discrete* machines because it is possible to describe their possible states as a discrete set, with the motion of the machines occurring by jumping from one state to another. Turing remarks that all machinery can be regarded as continuous (where the states form a continuous manifold and the behavior of the machine is described by a curve on this manifold) but "when it is possible to regard it as discrete it is usually best to do so. Moreover machineries are called "controlling" if they only deal with information, and "active" if aim at producing some definite physical effect. A bulldozer will be a continuous and active machine, a telephone continuous and controlling. But also brains can be considered machines and they are – Turing says "probably" – continuous and controlling but "very similar to much discrete machinery" (p. 5).



> Brains very nearly fall into this class [discrete controlling machinery – when it is natural to describe its possible states as a discrete set] and there seems every reason to believe that they could have been made to fall genuinely into it without any change in their essential properties. However, the property of being "discrete" is only an advantage for the theoretical investigator, and serves no evolutionary purpose, so we could not expect Nature to assist us by producing truly "discrete brains" (p. 6).

Brains can be treated as machines but they can also be considered discrete machines. The epistemological reason is clear: this is just an advantage for the "theoretical investigator" that aims at knowing what are

intelligent machines, but certainly it would not be an evolutionary advantage. "Real" humans brains are of course continuous systems, only "theoretically" they can be treated as discrete.
Following Turing's perspective we have derived two new achievements about machines and intelligence: brains can be considered machines, the simplest nervous systems with a random arrangement of neurons can be considered unorganized machines, in both cases with the property of being "discrete".

Mimicking Human Education

Turing adds:

> The types of machine that we have considered so far are mainly ones that are allowed to continue in their own way for indefinite periods without interference from outside. The universal machines were an exception to this, in that from time to time one might change the description of the machine which is being imitated. We shall now consider machines in which such interference is the rule rather than the exception (p. 11).

Screwdriver interference is when parts of the machine are removed and replaced with others, giving rise to completely new machines. *Paper* interference is when mere communication of information to the machine modifies its behavior. It is clear that in the case of the universal machine, paper interference can be as useful as screwdriver interference: we are interested in this kind of interference. We can say that each time an interference occurs the machine is probably changed. It has to be noted that paper interference provides information that is both "external" and "material" (further consideration on the status of this information are given below section 5.)
Turing thought that the fact that human beings have made machinery able to imitate any small part of a human being was positive in order to believe in the possibility of building thinking machinery: trivial examples are the microphone for the ear, and the television camera for the eye. What about the nervous system? We can copy the behavior of nerves with suitable electrical models and the electrical circuits which are used in electronic computing machinery seem to have essential properties of nerves because they are able to transmit information and to store it.
Education in human beings can model "education of machinery" "Mimicking education, we should hope to modify the machine until it could be relied on to produce definite reactions to certain commands" (p. 14). A graduate has had interactions with other human beings for twenty years or more and at the end of this period "a large number of standard routines will have been superimposed on the original pattern of his brain" (*ibid*.).
Turing maintains that
1) in human beings the interaction is manly with other human and the receiving of visual and other stimuli constitutes the main forms of interference;
2) it is only when a human being is "concentrating" that s/he approximates a machine without interference;

3) even when a human being is concentrating his behavior is mainly conditioned by previous interference.

Brains as Unorganized and Organized Machines

The Infant Cortex as an Unorganized Machine

In many unorganized machines when a configuration[20] is reached and possible interference suitably constrained, the machine behaves as one organized (and even universal) machine for a definite purpose. Turing provides the example of a B-type unorganized machine with sufficient units where we can find particular initial conditions able to make it a universal machine also endowed with a given storage capacity. The set up of these initial conditions is called "organizing the machine" that indeed is seen a kind of "modification" of a preexisting unorganized machine through external interference.

Infant brain can be considered an unorganized machine. Given the analogy previously established (cf. subsection 1.1 above "Logical, Practical, Unorganized, and Paper Machines), what are the events that modify it in an organized universal brain/machine? "The cortex of an infant is an unorganized machinery, which can be organized by suitable interference training. The organization might result in the modification of the machine into a universal machine or something like it. […] This picture of the cortex as an unorganized machinery is very satisfactory from the point of view of evolution and genetics." (p. 16). The presence of human cortex is not meaningful in itself: "[…] the possession of a human cortex (say) would be virtually useless if no attempt was made to organize it. Thus if a wolf by a mutation acquired a human cortex there is little reason to believe that he would have any selective advantage" (*ibid*.). Indeed the exploitation of a big cortex (that is its possible organization) requires a suitable environment: "If however the mutation occurred in a milieu where speech had developed (parrot-like wolves), and if the mutation by chance had well permeated a small community, then some selective advantage might be felt. It would then be possible to pass information on from generation to generation. (*ibid*.).

Hence, organizing human brains into universal machines strongly relates to the presence of

1) *speech* (even if only at the level rudimentary parrot-like wolves)

2) and a *social setting* where some "techniques" are learnt ("the isolated man does not develop any intellectual power. It is necessary for him to be immersed in an environment of other men, whose techniques he absorbs during the first twenty years of his life. He may then perhaps do a little research of his own and make a very few discoveries which are passed

20 A configuration is a state of a discrete machinery.

on to other men. From this point of view the search for new techniques must be regarded as carried out by human community as a whole, rather than by individuals" (p. 23).

This means that a big cortex can provide an evolutionary advantage only in presence of that massive storage of information and knowledge on external supports that only an already developed small community can possess. Turing himself consider this picture rather speculative but evidence from paleoanthropology can support it, as I will describe in the following section.

Moreover, the training of a human child depends on a system of rewards and punishments, that suggests that organization can occur only through two inputs. The example of an unorganized P-type machine, that can be regarded as a LCM without a tape and largely incompletely described, is given. Through suitable stimuli of pleasure and pain (and the provision of an external memory) the P-type machine can become an universal machine (p. 20).

When the infant brain is transformed in an intelligent one both discipline and initiative are acquired: "to convert a brain or machine into a universal machine is the extremest form of discipline. […] But discipline is certainly not enough in itself to produce intelligence. That which is required in addition we call initiative. […] Our task is to discover the nature of this residue as it occurs in man, and try and copy it in machines" (p. 21).

Examples of problems requiring initiative are the following: "Find a number *n* such that…", "see if you can find a way of calculating the function which will enable us to obtain the values for arguments….". The problem is equivalent to that of finding a program to put on the machine in question.

We have seen how a brain can be "organized", but how is the relation of that brain with the idea of "mimetic mind"?

From the Prehistoric Brains to the Universal Machines

I have said that a big cortex can provide an evolutionary advantage only in presence of a massive storage of information and knowledge on external supports that only an already developed small community of human beings can possess. Evidence from paleoanthropology seems to support this perspective. Some research in cognitive paleoanthropology teaches us that high level and reflective consciousness in terms of thoughts about our own thoughts and about our feelings (that is consciousness not merely considered as raw sensation) is intertwined with the development of *modern language* (speech) and *material culture*. After 250.000 years ago several hominid species had brains as large as ours today, but their behavior lacked any sign of art or symbolic behavior. If we consider high-level consciousness as related to a high-level organization – in Turing's sense – of human cortex, its origins can be

related to the active role of environmental, social, linguistic, and cultural aspects.

Handaxes were made by Early Humans and firstly appeared 1,4 million years ago, still made by some of the Neanderthals in Europe just 50.000 years ago. The making of handaxes is strictly intertwined with the development of consciousness. Many needed capabilities constitute a part of an evolved psychology that appeared long before the first handaxed were manufactured. Consequently, it seems humans were pre-adapted for some components required to make handaxes (Mithen, 1996, 1999):

1. imposition of *symmetry* (already evolved through predators escape and social interaction). It has been an unintentional by-product of the bifacial knapping technique but also deliberately imposed in other cases. It is also well-known that the attention to symmetry may have developed through social interaction and predator escape, as it may allow one to recognize that one is being directly stared at (Dennett, 1991). It seems that “Hominid handaxes makers may have been keying into this attraction to symmetry when producing tools to attract the attention of other hominids, especially those of the opposite sex” (Mithen, 1999, p. 287);
2. understanding *fracture dynamics* (for example evident from Oldowan tools and from nut cracking by chimpanzees today);
3. ability to *plan* ahead (modifying plans and reacting to contingencies, such unexpected flaws in the material and miss-hits), still evident in the minds of Oldowan tool makers and in chimpanzees;
4. high degree of *sensory-motor control*: The origin of this capability is usually tracked back to encephalization – the increased number of nerve tracts and of the integration between them allows for the firing of smaller muscle groups - and bipedalism – that requires a more complex integrated highly fractionated nervous system, which in turn presupposes a larger brain.

The combination of these four resources produced the birth of what Mithen calls *technical intelligence* of early human mind, that is consequently related to the construction of handaxes. Indeed they indicate high intelligence and good health. They cannot be compared to the artefacts made by animals, like honeycomb or spider web, deriving from the iteration of fixed actions which do not require consciousness and intelligence.

Private Speech and Fleeting Consciousness

Two central factors play a fundamental role in the combination of the four resources above:

- the exploitation of *private speech* (speaking to oneself) to trail between planning, fracture dynamic, motor control and symmetry

(also in children there is a kind of private muttering which makes explicit what is implicit in the various abilities);
- a good degree of *fleeting consciousness* (thoughts about thoughts).

In the meantime these two aspects played a fundamental role in the development of consciousness and thought:

> So my argument is that when our ancestors made handaxes there were private mutterings accompanying the crack of stone against stone. Those private mutterings were instrumental in pulling the knowledge required for handaxes manufacture into an emergent consciousness. But what type of consciousness? I think probably one that was fleeting one: one that existed during the act of manufacture and that did not the endure. One quite unlike the consciousness about one's emotions, feelings, and desires that were associated with the social world and that probably were part of a completely separated cognitive domain, that of social intelligence, in the early human mind" (p. 288).

This use of private speech can be certainly considered a "tool" for organizing brains and so for manipulating, expanding, and exploring minds, a tool that probably evolved with another: talking to each other. Both private and public language act as tools for thought and play a central role in the evolution of consciousness.

Material Culture and Semiosis

Another semiotic tool appeared in the latter stages of human evolution, that played a great role in the evolutions of minds in mimetic minds, that is in a further organization of human brains. Handaxes are at the birth of *material culture*, so as new cognitive chances can co-evolve:

- the mind of some early humans, like the Neanderthals, were constituted by relatively isolated cognitive domains, Mithen calls *different intelligences*, probably endowed with different degree of consciousness about the thoughts and knowledge within each domain (natural history intelligence, technical intelligence, social intelligence). These isolated cognitive domains became integrated also taking advantage of the role of public language;
- *degrees of high level consciousness* appear, human beings need thoughts about thoughts;
- *social intelligence* and *public language* arise.

It is extremely important to stress that *material culture* is not just the product of this massive cognitive chance but also cause of it. "The clever trick that humans learnt was to *disembody* their minds into the material world around them: a linguistic utterance might be considered as a disembodied thought. But such utterances last just for a few seconds. Material culture endures" (p. 291).
In this perspective we acknowledge that material artefacts are tools for thoughts as is language: tools for exploring, expanding, and manipulating

our own minds. In this regard the evolution of culture is inextricably linked with the evolution of consciousness and thought.

Early human brain becomes a kind of universal “intelligent” machine, extremely flexible so that we did no longer need different “separated” intelligent machines doing different jobs. A single one will suffice. As the engineering problem of producing various machines for various jobs is replaced by the office work of “programming” the universal machine to do these jobs, so the different intelligences become integrated in a new universal device endowed with a high-level type of consciousness.

From this perspective the expansion of the minds is in the meantime a continuous process of *disembodiment* of the minds themselves into the *material world* around them. In this regard the evolution of the mind is inextricably linked with the evolution of large, integrated, material cognitive systems. In the following sections I will illustrate this extraordinary interplay between human brains and the cognitive systems they make, which is at the origins of the first interesting features of the modern human mind. What I call *semiotic brains* are brains that make up a series of signs and that are engaged in making or manifesting or reacting to a series of signs: through this semiotic activity they are at the same time engaged in “being minds” and so in thinking intelligently. An important effect of this semiotic activity of brains is a continuous process of disembodiment of mind that exhibits a new cognitive perspective on the mechanisms underling the semiotic emergence of meaning processes (cf. the following section).

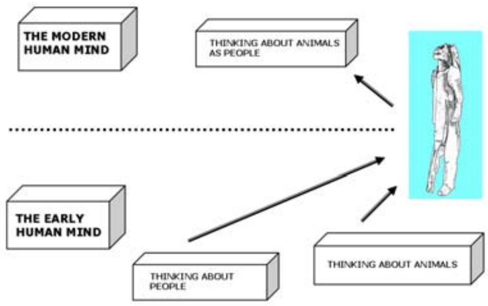
Semiotic Delegations through the Disembodiment of Mind

A wonderful example of the cognitive effects of the disembodiment of mind is the carving of what most likely is the mythical being from the last ice age, 30.000 years ago, a half human/half lion figure carved from mammoth ivory found at Hohlenstein Stadel, Germany.

> An evolved mind is unlikely to have a *natural home* for this being, as such entities do not exist in the natural world: so whereas evolved minds could think about humans by exploiting modules shaped by natural selection, and about lions by deploying content rich mental modules moulded by natural selection and about other lions by using other content rich modules from the natural history cognitive domain, how could one think about entities that were part human and part animal? Such entities had no home in the mind (p. 291).

A mind consisting of different “separated intelligences” (for instance “thinking about animals” as separated from “thinking about people”) cannot come up with such entity. The only way is *to extend* the mind into the *material word*, exploiting rocks, blackboards, paper, ivory, and writing, painting, and carving: “artefacts such as this figure play the role of anchors for ideas and have no *natural home* within the mind; for ideas that take us beyond those that natural selection could enable us to possess” (p. 291) (cf. Figure 2).

In the case of our figure we face with an anthropomorphic thinking created by the material representation serving to anchor the cognitive representation of supernatural being. In this case the material culture



disembodies thoughts, that otherwise will soon disappear, without being transmitted to other human beings. The early human mind possessed two separated intelligences for thinking about animals and people. Through the mediation of the material culture the modern human mind can *creatively* arrive to "internally" think about animals and people at the same time.

Figure 2.

Artefacts as *external semiotic objects* allowed humans to loosen and cut those chains on our unorganized brains imposed by our evolutionary past. Chains that always limited the brains of other human beings, such as the Neanderthals. Loosing chains and securing ideas to external objects was a way to re-organize brains as universal machines for thinking. Still important in human reasoning and in computational modeling is the role of external representations and mediators. I devoted part of my research to illustrate their role at the epistemological and ethical level (Magnani, 2001, 2005).
In the remaining part of the paper I will describe the centrality to semiotic cognitive information processes of the disembodiment of mind from the point of view of the cognitive interplay between internal and external representations. I consider this interplay critical in analyzing the relation between meaningful semiotic internal resources and devices and their dynamical interactions with the externalized semiotic materiality already stocked in the environment. Hence, minds are "extended" and artificial in themselves.

Mimetic and Creative Representations

We have seen that unorganized brains organize themselves through a semiotic activity that is reified in the external environment and then re-projected and reinterpreted through new configurations of neural

networks and chemical processes. I also think the disembodiment of mind can nicely account for low-level semiotic processes of meaning creation, bringing up the question of how could higher-level processes be comprised and how would they interact with lower-level ones.

External and Internal Representations

We have said that through the mediation of the material culture the modern human mind can creatively arrive to internally think about animals and people at the same time. We can also account for this process of disembodiment from an interesting cognitive point of view.
I maintain that representations can be external and internal. We can say that
– *external representations* are formed by external materials that express (through reification) concepts and problems that do not have a natural home in the brain;
– *internalized representations* are internal re-projections, a kind of recapitulations, (learning) of external representations in terms of neural patterns of activation in the brain. They can sometimes be "internally" manipulated like external objects and can originate new internal reconstructed representations through the neural activity of *transformation* and *integration*.
This process explains why human beings seem to perform both computations of a *connectionist* type such as the ones involving representations as
– (I Level) *patterns of neural activation* that arise as the result of the interaction between body and environment – that interaction that is extremely fruitful for creative results - (and suitably shaped by the evolution and the individual history): pattern completion or image recognition,
and computations that use representations as
– (II Level) *derived combinatorial syntax and semantics* dynamically shaped by the various external representations and reasoning devices found or constructed in the environment (for example geometrical diagrams in mathematical creativity); they are neurologically represented contingently as pattern of neural activations that "sometimes" tend to become stabilized structures and to fix and so *to permanently belong to the I Level* above.
The I Level originates those *sensations* [they constitute a kind of "face" we think the world has], that provide room for the II Level to reflect the structure of the environment, and, most important, that can follow the computations suggested by these external structures. It is clear we can now conclude that the growth of the brain and especially the synaptic and dendritic growth are profoundly determined by the environment.
When the fixation is reached the patterns of neural activation no longer need a direct stimulus from the environment for their construction. In a certain sense they can be viewed as *fixed internal records* of *external*

structures that *can exist* also in the absence of such external structures. These patterns of neural activation that constitute the I Level Representations always keep record of the experience that generated them and, thus, always carry the II Level Representation associated to them, even if in a different form, the form of *memory* and not the form of a vivid sensorial experience. Now, the human agent, via neural mechanisms, can retrieve these II Level Representations and use them as *internal* representations or use parts of them to construct new internal representations very different from the ones stored in memory (cf. also Gatti and Magnani, 2005).

Human beings delegate cognitive features to external representations because in many problem solving situations the internal computation would be impossible or it would involve a very great effort because of human mind's limited capacity. First a kind of alienation is performed, second a recapitulation is accomplished at the neuronal level by re-representing internally that which was "discovered" outside. Consequently only later on we perform cognitive operations on the structure of data that synaptic patterns have "picked up" in an analogical way from the environment. Internal representations used in cognitive processes have a deep origin in the experience lived in the environment.

I think there are two kinds of artefacts that play the role of *external objects* (representations) active in this process of disembodiment of the mind: *creative* and *mimetic*. Mimetic external representations mirror concepts and problems that are already represented in the brain and need to be enhanced, solved, further complicated, etc., so they sometimes can give rise to new concepts, models, and perspectives.[21]

Following my perspective it is at this point evident that the "mind" transcends the boundary of the individual and includes parts of that individual's environment. In the following section I will illustrate some fundamental aspects of the interplay above in the light of basic semiotic aspects of inferences in a Peircian perspective.

Inferences and Cognitive Semiosis beyond Peirce

Peirce stated that all thinking is in signs, and signs can be icons, indices, or symbols. Moreover, all *inference* is a form of sign activity, where the word sign includes "feeling, image, conception, and other representation" (Peirce, *CP*, 5.283), and, in Kantian words, all synthetic forms of cognition. That is, a considerable part of the creative meaning processes is *model-based*. Moreover, a considerable part of the meaningful behavior (not only in science) occurs in the middle of a relationship

21 I studied the role of diagrams in mathematical reasoning endowed both of mirroring and creative roles (Magnani and Dossena, 2003). I also think this discussion about external and internal representations can be used to enhance the Representational Redescription model introduced by Karmiloff-Smith (1992), that accounts for how these levels of representation are generated in the infant mind.

between brains and external objects and tools that have received cognitive and/or epistemological delegations (cf. the previous and the following subsection).

Following this Peircian perspective about inference I think it is extremely useful from a cognitive point of view to consider the concept of reasoning in a very broad way (cf. also Brent, 2000, p. 8). We have three cases:

1) reasoning can be fully conscious and typical of high-level worked-out ways of inferring, like in the case of scientists' and professionals' performances;

2) reasoning can be "acritical" (Peirce, *CP*, 5.108), which includes every day inferences in conversation and in various ordinary patterns of thinking;

3) reasoning can resort to "operations of the mind which are logically analogous to inference excepting only that they are unconscious and therefore uncontrollable and therefore not subject to logical criticism" (Peirce, *CP*, 5.108).

Immediately Peirce adds a note to the third case "But that makes all the difference in the world; for inference is essentially deliberate, and self-controlled. Any operation which cannot be controlled, any conclusion which is not abandoned, not merely as soon as criticism has pronounced against it, but in the very act of pronouncing that decree, is not of the nature of rational inference – is not reasoning" (*ibid.*).

As Colapietro clearly states (2000, p. 140), it seems that for Peirce human beings semiotically involve unwitting trials and unconscious processes. Moreover, it seems clear that unconscious thought can be in some sense considered "inference", even if not rational; indeed, Peirce says, it is not reasoning. Peirce further indicates that there are in human beings multiple trains of thought at once but only a small fraction of them is conscious, nevertheless the prominence in consciousness of one train of thought is not to be interpreted an interruption of other ones.

In this Peircian perspective, which I adopt in this essay, where inferential aspects of thinking dominate, there is no intuition, in an anti-Cartesian way. We know all important facts about ourselves in an *inferential* abductive way:

> […] we first form a definite idea of ourselves as a hypothesis to provide a place in which our errors and other people's perceptions of us can happen. Furthermore, this hypothesis is constructed from our knowledge of "outward" physical facts, such things as the sounds we speak and the bodily movements we make, that Peirce call signs (cit., p. 10).

Recognizing in a series of *material*, physical events, that they make up a series of signs, is to know the existence of a *mind* (or of a group of minds) and to be absorbed in making, manifesting, or reacting to a series of signs is to be absorbed in "being a mind". "[…] all thinking is dialogic in form" (Peirce, *CP*, 6.338)**,** both at the intrasubjective[22] and intersubjective

22 "One's thoughts are what he is 'saying to himself', that is saying to that other self that is just coming to life in the flow of time. When one reasons, it that critical self that one is trying to persuade: and all thought whatsoever is a sign, and is mostly in the nature of language" (Peirce, CP, 5.421).

level, so that we see ourselves exactly as others see us, or see them exactly as they see themselves, and we see ourselves through our own speech and other interpretable behaviors, just others see us and themselves in the same way, in the commonality of the whole process (Brent, 2000, p. 10).

In this perspective minds are material like brains, in so far as they consist in intertwined internal and external semiotic processes: "[…] the psychologists undertake to locate various mental powers in the brain; and above all consider it as quite certain that the faculty of language resides in a certain lobe; but I believe it comes decidedly nearer the truth (though not really true) that language resides in the tongue. In my opinion it is much more true that the thoughts of a living writer are in any printed copy of his book than they are in his brain" (Peirce, *CP*, 7.364).

Man is an External Sign

Peirce's semiotic motto "man is an external sign" is very clear about the materiality of mind and about the fact that the conscious self[23] is a cluster actively embodied of flowing intelligible signs:

> It is sufficient to say that there is no element whatever of man's consciousness which has not something corresponding to it in the word; and the reason is obvious. It is that the word or sign which man uses *is* the man himself. For, as the fact that every thought is a sign, taken in conjunction with the fact that life is a train of thoughts, proves that man is a sign; so, that every thought is an *external sign*, proves that man is an external sign. That is to say, the man and the *external sign* are identical, in the same *sense* in which the words *homo* and *man* are identical. Thus my language is the sum total of myself; for the man is the thought (Peirce, *CP*, 5.314).

It is by way of signs that we ourselves *are* semiotic processes – for example a more or less coherent cluster of narratives. If all thinking is in signs it is not true that thoughts are in us because we are in thoughts: "[…] man is a sign developing according to the laws of inference. […] the entire phenomenal manifestation of mind is a sign resulting from inference" (Peirce, *CP*, 5.312 and 5.313).

Moreover, the "person-sign" is future-conditional, that is not fully formed in the present but depending on the future destiny of the concrete semiotic activity (future thoughts and experience of the community) in which s/he will be involved. If Peirce maintains that when we think we appear as a sign (Peirce, *CP*, 5.283) and, moreover, that everything is present to us is a phenomenal manifestation of ourselves, feelings, images, diagrams, conceptions, schemata, and other representations are phenomenal manifestations that become available for interpretations and thus are guiding our actions in a positive or negative way. They become *signs* when we think and interpret them. It is well-known that for Peirce all semiotic experience – and thus abduction - is also providing a guide for

23 Consciousness arises as "a sort of public spirit among the nerve cells" (Peirce, CP, 1.354).

action. Indeed the whole function of thought is to produce habits of action.[24]

Mimetic Minds

It is well-known that there are external representations that are representations of other external representations. In some cases they carry new scientific knowledge. To make an example, Hilbert's *Grundlagen der Geometrie* is a "formal" representation of the geometrical problem solving through diagrams: in Hilbertian systems solutions of problems become proofs of theorems in terms of an axiomatic model. In turn a calculator is able to re-represent (through an artifact) (and to perform) those geometrical proofs with diagrams already performed by human beings with pencil and paper. In this case we have representations that *mimic* particular cognitive performances that we usually attribute to our *minds*.

We have seen that our brains delegate cognitive (and epistemic) roles to externalities and then tend to "adopt" and recapitulate what they have checked occurring outside, over there, after having manipulated – often with creative results - the external invented structured model. A simple example: it is relatively neurologically easy to perform an addition of numbers by depicting in our *mind* – thanks to that brain device that is called visual buffer – the images of that addition *thought* as it occurs concretely, with paper and pencil, taking advantage of external materials. We have said that mind representations are also over there, in the environment, where mind has objectified itself in various structures that *mimic* and *enhance* its internal representations.

Turing adds a new structure to this list of external objectified devices: an abstract tool (LCM) endowed with powerful mimetic properties. We have concluded the previous section remarking that the "mind" is in itself extended and, so to say, both internal and external: the mind transcends the boundary of the individual and includes parts of that individual's environment. Turing's LCM, which is an externalized device, is able to mimic human cognitive operations that occur in that interplay between the internal mind and the external one. Indeed Turing already in 1950 maintains that, taking advantage of the existence of the LCM, "Digital computers […] can be constructed, and indeed have been constructed, and […] they can in fact mimic the actions of a human computer very closely" (Turing, 1950).

In the light of my perspective both (Universal) Logical Computing Machine (LCM) (the theoretical artifact) and (Universal) Practical Computing Machine (PCM) (the practical artifact) are *mimetic minds* because they are able to mimic the mind in a kind of universal way

---

24 Cf. for example the contributions contained in recent special issue of the journal Semiotica 153 (1/4) devoted to abduction and edited by Queiroz and Merrell.

(wonderfully continuing the activity of disembodiment of minds our ancestors rudimentary started). LCM and PCM are able to re-represent and perform in a very powerful way plenty of cognitive skills of human beings.
Universal Turing machines are discrete-state machines, DMS, "with a Laplacian behavior" (Longo, 2002; Lassègue, 1998, 2002): "it is always possible to predict all future states") and they are equivalent to all formalisms for computability (what is thinkable is calculable and mechanizable), and because universal they are able to simulate – that is to *mimic* – any human cognitive function, that is what is usually called mind.
Universal Turing machines are just a further extremely fruitful step of the *disembodiment* of the mind I have described above. A natural consequence of this perspective is that they do not represent (against classical AI and modern cognitivist computationalism) a "knowledge" of the mind and of human intelligence. Turing is perfectly aware of the fact that brain is not a DSM, but as he says, a "continuous" system, where instead a mathematical modeling can guarantee a satisfactory scientific intelligibility (cf. his studies on morphogenesis).
We have seen that our brains delegate cognitive (and epistemic) roles to externalities and then tend to "adopt" what they have checked occurring outside, over there, in the external invented structured and model.
Our view about the disembodiment of the mind certainly involves that the Mind/Body dualist view is less credible as well as Cartesian computationalism. Also the view that Mind is Computational independently of the physical (functionalism) is jeopardized. In my perspective on human cognition in terms of mimetic minds we no longer need Descartes dualism: we only have brains that make up large, integrated, material cognitive systems like for example LCMs and PCMs. The only problem seems "How meat knows":we can reverse the Cartesian motto and say "sum ergo cogito". In this perspective what we usually call mind simply consist sin the union of both the changing neural configurations of brains together with those large, integrated, and material cognitive systems the brains themselves are continuously building.

Conclusion

The main thesis of this paper is that the disembodiment of mind is a significant cognitive perspective able to unveil some basic features of creative thinking. Its fertility in explaining the interplay between internal and external levels of cognition is evident. I maintain that various aspects of cognition could take advantage of the research on this interplay: for instance study on external mediators can provide a better understanding of the processes of explanation and discovery in science and in some areas of artificial intelligence related to mechanizing discovery

processes.[25] For example, concrete manipulations of external objects influence the generation of hypotheses: what I have called manipulative abduction shows how we can find methods of constructivity in scientific and everyday reasoning based on external models and "epistemic mediators" (Magnani, 2004).

Finally, I think the cognitive role of what I call "mimetic minds" can be further studied also taking advantage of the research on hypercomputation. The imminent construction of new types of universal "abstract" and "practical" machines will constitute important and interesting new "mimetic minds" externalized and available over there, in the environment, as sources of mechanisms underlying the emergence of new meaning processes. They will provide new tools for creating meaning formation in classical areas like analogical, visual, and spatial inferences, both in science and everyday situations, so that this can extend the epistemological and the psychological theory.

References

Brent, J. (2000). A brief introduction to the life and thought of Charles Sanders Peirce. In J. Muller & J. Brent (Eds.), *Peirce, Semiosis, and Psychoanalysis*. Baltimore and London: John Hopkins (pp. 1-14).

Colapietro, V. (2000). Futher consequences of a singular capacity. In J. Muller & J. Brent (Eds.) *Peirce, Semiosis, and Psychoanalysis*. Baltimore and London: John Hopkins (pp. 136-58).

Dennett, D. (1991). Consciousness Explained. Little, Brown, and Company, New York, 1991.

Gatti, A and Magnani, L. (2005) On the representational role of the environment and on the cognitive nature of manipulations, in: L. Magnani and R. Dossena (eds.), *Computing, Philosophy, and Cognition*, College Publications, London (Proceedings of the International Conference E-CAP2004), pp. 227-242.

Hameroff, S R., Kaszniak, A. W. and Chalmers, D. J. (Eds.) (1999). *Toward a Science of Consciousness III. The Third Tucson Discussions and Debates*. MIT Press, Cambridge, MA, 1999.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT Press, Cambridge, MA, 1992.

Lassègue, J. (1998). *Turing*. Les Belles Lettres, Paris, 1998.

Lassègue, J. (2002). Turing entre formel et forme; remarque sur la convergence des perspectives morphologiques. *Intellectica*, 35(2):185-198, 2002.

Longo, G. (2002). Laplace, Turing, et la géométrie impossible du "jeu de l'imitation": aléas, determinisme e programmes dans le test de Turing. *Intellectica,* 35(2):131-161, 2002.

Magnani, L. (2001). *Abduction, Reason, and Science. Processes of Discovery and Explanation*. Kluwer Academic/Plenum Publishers, New York, 2001.

Magnani, L. (2004). Conjectures and manipulations. Computational modeling and the extra-theoretical dimension of scientific discovery. *Minds and Machines* 14: 507-537, 2004.

Magnani, L. (2007). *Knowledge as a Duty. Distributed Morality in a Technological World*, Cambridge University Press, Cambridge, forthcoming.

Magnani, L. and Dossena, R. (2003). Perceiving the infinite and the infinitesimal world: unveiling and optical diagrams and the construction of mathematical concepts. In *Proceedings of CogSci2003*. CD-ROM produced by X-CD Technologies, Boston, MA, 2003.

Magnani, L. Nersessian, N. J., and Pizzi, C. (Eds.) (2002). *Logical and Computational Aspects of Model-Based Reasoning*, Kluwer Academic, Dordrecht, 2002.

25 On the recent achievements in the area of the machine discovery simulations of model-based creative tasks cf. [Magnani, Nersessian, and Pizzi, 2002].

Mithen, S (1996). *The Prehistory of the Mind. A Search for the Origins of Art, Religion, and Science*. Thames and Hudson, London, 1996.

Mithen, S. (1999). Handaxes and ice age carvings: hard evidence for the evolution of consciousness. In (Hameroff, et al., 1999), pp. 281-296.

Peirce, C.S. (1931-1958) (*CP*). *Collected Papers*, 8 vols. C. Hartshorne & P. Weiss (vols. I-VI), (Eds.), & A.W. Burks (vols. VII-VIII) (Ed.), Cambridge, MA: Harvard University Press.

Turing, A. M. (1937). On computable numbers with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230-265, 1937.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 49:433-460, 1950. Also in (Turing, 1992a), pp. 133-160.

Turing, A. M. (1969), Intelligent machinery [1948]. In Bernard Meltzer and Donald Michie (Eds.), *Machine Intelligence* 5:3-23. Also in (Turing, 1992a), pp. 3-23.

Turing. A. M. (1992a). *Collected Works of Alan Turing, Mechanical Intelligence*. Ed. by Darrel C. Ince, Elsevier, Amsterdam, 1992.

Turing, A. M. (1992b), *Collected Works of Alan Turing, Morphogenesis*, ed. by Peter T. Saunders, Elsevier, Amsterdam.

BELIEF AND DESIRE: A LOGICAL ANALYSIS

DANIEL VANDERVEKEN

Université du Québec

www.vanderveken.org[26]

Propositional attitudes directed at facts of the world are cognitive or volitive. Cognitive attitudes contain beliefs and volitive attitudes desires. Beliefs are satisfied whenever they are true and desires whenever they are realized. The main purpose of this paper is to analyze the logical form of such primitive kinds of attitudes and to explicate their conditions of possession and of satisfaction. According to standard epistemic logic human agents are either perfectly rational or totally irrational. I will advocate an intermediate position compatible with philosophy of mind and psychology according to which agents are minimally rational. They can be inconsistent but they never are entirely irrational. In order to account for the imperfect but minimal rationality of human agents I will explicate both *subjective* and *objective possibilities* in the logic of attitudes. For that purpose, I will exploit the resources of a non classical propositional *predicative logic* that distinguishes propositions with the same truth conditions that do not have the same cognitive value. In my approach, the relations of compatibility with the satisfaction of agents' beliefs and desires are also explicated in a finer way so as to avoid current epistemic and volitive paradoxes. I will use both proof- and model-theory in order to formulate my logic of belief and desire. At the end I will enumerate important valid laws governing such attitudes.

As Descartes pointed out in his treatise on ***Les passions de l'âme***[27]**,** beliefs and desires are primitive kinds of attitudes. Every propositional attitude is cognitive or volitive. All cognitive attitudes (e.g. conviction, faith, confidence, knowledge, certainty, presumption, pride, arrogance, surprise, amazement, stupefaction, prevision, anticipation, expectation) contain beliefs. Similarly, all volitive attitudes (e.g. wish, will,

---

26 I am grateful to Hugolin Bergier, David Kaplan, Kenneth MacQueen and John Searle for critical remarks on a first draft of this paper.

27 The treatise Les passions de l'âme is reedited in R. Descartes Œuvres complètes La Pléiade Gallimard 1953

intention, ambition, project, hope, aspiration, satisfaction, pleasure, enjoyment, delight, gladness, joy, elation, amusement, fear, regret, sadness, sorrow, grief, remorse, terror) contain desires. So beliefs and desires are essential components of all propositional attitudes. The main purpose of this chapter is to analyze their logical form.

Beliefs and desires have *intentionality*: they are *directed at* objects and facts of the world. So these attitudes have *conditions* of *possession* and of *satisfaction* that are logically related but different. In order that an agent *possesses* a belief or a desire, he must be in a certain mental state. Whoever *possesses* a *belief* represents how things are in the world according to him. Whoever *possesses* a *desire* represents how he prefers things to be in the world. In order that a belief or desire is *satisfied*, represented things must be or turn out to be in the world as the agent represents them. My main objective here is to contribute to the foundations of the logic of attitudes in formulating a general *recursive theory of conditions of possession and of satisfaction of beliefs and desires*, the two primitive kinds of *attitudes*. In my view, more complex psychological modes such as knowledge, certainty, will and intention divide into other components than the *basic traditional categories of cognition* and *volition*. Complex modes have a *proper way* of believing or desiring, proper *conditions on their propositional content* or proper *preparatory conditions*. I have formulated a recursive definition of all psychological modes and analyzed complex attitudes in another more general paper.[28]

In the first section I will consider basic problems of the standard logic of propositional attitudes. Like many philosophers of mind, I do not think that human agents are either perfectly rational or totally irrational. On the contrary, I advocate an intermediate position: agents are minimally rational. They can be inconsistent but they never are entirely inconsistent. They do not make all logical inferences but they always make many. We need to consider *subjective* as well as *objective possibilities* in the logic of attitudes and action in order to account for the imperfect but minimal rationality of human agents. For that purpose, I will present in the second section a *predicative logic* that provides a much finer criterion of propositional identity than standard propositional logic. In the third section I will proceed to the analysis of the general categories of cognition and volition. In my approach, the relations of compatibility with the satisfaction of agents' beliefs and desires are explicated in a finer way. Current epistemic and volitive paradoxes are eliminated. I will formulate the ideography and formal semantic of my logic of belief and desire in the fourth section. In the fifth section I will formulate an axiomatic system and in the last section I will enumerate important valid laws governing beliefs and desires.

---

28 D. Vanderveken "Fondements de la logique des attitudes" in press in the issue Language and Thought of Manuscrito. The present paper is based on the theory developed in the first part of that French paper.

1. Basic problems of the standard logic of attitudes

Following Carnap[29], standard propositional logic tends to identify propositions that have the same truth values in the same possible circumstances. However it is clear that strictly equivalent propositions are not the contents of the same attitudes and intentional actions just as they are not the senses of synonymous sentences. We absolutely need a much finer criterion of propositional identity than strict equivalence in logic for the purposes of philosophy of mind, action and language. Maybe Carnap's reduction of Fregean *senses* to *intensions* enables us to define logical truth in the special cases of modal and temporal logics? But it does not work for richer logics dealing with attitudes, actions and illocutions. We need a better logic of sense in order to formulate an adequate theory of meaning, action and thought.

We need first to analyze the logical form of propositions so as to distinguish propositions with the same truth conditions that do not have the same cognitive value. Clearly we do not know *a priori* by virtue of competence the necessary truth of many propositions. We have to learn a lot of essential properties of objects to which we refer. By *essential property* of an object I mean here a property that it really possesses in any possible circumstance.[30] We discovered in modern times that whales are essentially mammals. So we can ignore necessary truths. We can even be inconsistent and believe necessary falsehoods. (We believed in the past that whales are fishes). However we always remain paraconsistent. As the Greek philosophers already pointed out, we cannot believe that every proposition is true (the sophist's paradox) or false (the skeptic's paradox).. Any adequate logic of attitudes and action has to account for such facts. Few necessarily true propositions are pure *tautologies* such as the proposition that whales are whales that we know *a priori* to be true. What is the logical nature of such *tautologies*? My predicative propositional logic gives an answer to that question. It also explains why certain strictly equivalent propositions have a different cognitive value. This solves the first problem of propositional identity.

A second important problem of the standard logic of attitudes is related to the way in which it analyzes the *relations of compatibility with the truth of beliefs and the realization of desires of agents*. According to standard logic, such relations of psychological compatibility are simple modal relations of accessibility existing between agents and moments, on one hand, and possible circumstances, on the other hand. Thus in Hintikka's epistemic logic[31], possible circumstances are compatible with the truth of beliefs of agents at each moment of time. To each agent *a* and moment *m* there corresponds in each model a unique set *Belief*(*a,m*) of possible circumstances that are compatible with the truth of all beliefs that agent *a* has at moment *m*. Moreover, according to the standard

---

29 R. Carnap Meaning and Necessity University of Chicago Press 1956.

30 See A. Plantinga The Nature of Necessity Oxford University Press 1974.

31 J. Hintikka "Semantics for Propositional attitudes" in L. Linsky (ed) Reference and Modality Oxford U Press 1962

meaning postulate for belief propositions: an agent *a believes a proposition at a moment m* when that proposition is true in all possible circumstances belonging to the set *Belief (a,m)* of circumstances compatible with what that agent then believes.

Given such a formal approach, all human agents are either *perfectly rational* or *totally irrational*. On one hand, they believe all necessarily true propositions. And their beliefs are closed under strict implication. Whoever believes a proposition *eo ipso* believes all propositions that are strictly implied by that proposition. So human agents are perfectly rational when at least one possible circumstance is compatible with what they believe. Otherwise, they are totally irrational. Whoever believes a necessary falsehood *eo ipso* believes all propositions. However, all this is absolutely false according to standard philosophy of mind and empirical psychology. First of all, human agents are *not logically omniscient*. They ignore most logical truths and they do not draw all logical inferences. Moreover even when they are inconsistent, they never are entirely inconsistent. Problems are worse in the case of the logic of desire if we proceed according to the same approach. In that case, to each agent *a* and moment *m* there corresponds in each model a unique set *Desire*(*a,m*) of possible circumstances that are compatible with the satisfaction of all desires of that agent at that moment. We can make mistakes and wrongly believe necessarily false propositions. However, when we recognize their falsehood, we immediately stop believing them. On the contrary, it is not enough to learn that something is impossible in order to stop desiring it. We keep many desires that we know to be unrealizable. Moreover we never desire everything.

Some have advocated the introduction in epistemic logic of so-called *impossible circumstances* where necessarily false propositions would be true (where, for example, whales would be fishes). However, such a theoretical move is very *ad hoc*. Moreover it is neither necessary nor sufficient. So I prefer to keep only possible circumstances in models while changing the approach. In logic, *possible circumstances* are *objective possibilities* as Belnap says. So objects keep their essential properties (whales are really mammals) and necessarily false propositions remain false in all possible circumstances. However, according to human agents certain necessarily false propositions are true. We did believe that whales are fishes. This is an *epistemic possibility*. So we need to consider *subjective* in addition to *objective possibilities* in logic. Many subjective possibilities are not objective. There is no way to explicate pure subjective possibilities and to define adequately the notion of truth according to an agent within the standard approach. I will enrich the conceptual apparatus of propositional logic so as to analyze the logical form of necessarily false propositions that we can believe and desire. I will also provide a better analysis of the compatibility relation with respect to the satisfaction of attitudes of agents and adopt a finer meaning postulate for defining truth conditions of propositions attributing attitudes.

2. New principles of my predicative propositional logic

My propositional logic is *predicative* in the very general sense that it analyzes the logical form of propositions by taking into account predications that we make in expressing and understanding them.[32]

- In my view, each proposition has a finite *structure of constituents*. It predicates a positive number of *attributes* (properties or relations) of *objects subsumed under concepts*. Each proposition serves to make finitely many predications. As Frege and Russell pointed out, we understand a proposition when we understand which attributes objects must possess in a possible circumstance in order that this proposition be true in that circumstance.
- In addition to taking into account the structure of constituents of propositions, we also need a *better explication of their truth conditions*. We ignore in which possible circumstances most propositions are true because we ignore *real denotations* of most attributes and concepts in many possible circumstances. One can refer to Smith's murderer without knowing who he is. However we can always in principle think of persons who could be that murderer. Sometimes there are even suspects. So in any possible use and interpretation of language, there are a lot of *possible denotation assignments to attributes and concepts* in addition to the standard *real denotation assignment* which associates with each propositional constituent its actual denotation in every possible circumstance. They are functions of the same type. They, for example, associate with individual concepts a unique individual or no individual at all in each possible circumstance. According to a possible denotation assignment, Smith's wife murdered Smith. According to another possible assignment, another person is Smith's murderer. According to others, no one murdered Smith. By hypothesis, all possible denotation assignments respect *meaning postulates*. A murderer is not only an individual object; it is a person who has caused death.

We ignore the value of the real denotation assignment for most concepts and attributes in many possible circumstances. But we can in principle think of denotations that they could have. Moreover, when we have in mind certain concepts and attributes only some possible denotation assignments to these senses *are then compatible with* our beliefs. Persons born after Smith's death could not have murdered him. Suppose that according to the chief of police at the beginning of his investigation Smith's murderer is a foreigner. In that case, only possible denotation assignments according to which a foreigner falls under the concept of being Smith's murderer are then compatible with his actual beliefs. So in my approach, possible denotation assignments rather than possible circumstances are compatible with the beliefs of agents in possible circumstances. This is my way to account for subjective possibilities.

---

32 For more information on predicative logic see my collective book Logic Thought & Action Springer, 2005

- In predicative propositional logic, the *truth definition* is then relative to both possible circumstances and possible denotation assignments. Propositions are *true (or false) in a circumstance according to certain possible denotation assignments* to their constituents. In understanding propositions we in general do not know whether they are true or false. We just know that their truth in a possible circumstance is compatible with certain possible denotation assignments to their attributes and concepts, and incompatible with all others. Thus an elementary proposition predicating an extensional property of an object under a concept (e.g. the proposition that Smith's murderer is tall) is true in a possible circumstance according to a possible denotation assignment if and only if according to that assignment the person who falls under that concept has that property in that circumstance. Otherwise, that proposition is false in that very circumstance according to that assignment. We only know this by virtue of competence. Most propositions have therefore *a lot of possible* in addition to their real *truth conditions*. Suppose that a proposition is true according to a certain possible denotation assignment to its constituents in a certain set of possible circumstances. By definition, that proposition would be true in all and only these possible circumstances if that denotation assignment were the real one.

Of course, in order to be *true in a circumstance* a proposition has to be *true in that circumstance according to the real denotation assignment.* So among all possible truth conditions of a proposition, there are its *real Carnapian truth conditions* which correspond to the set of possible circumstances where it is true according to the real denotation assignment.

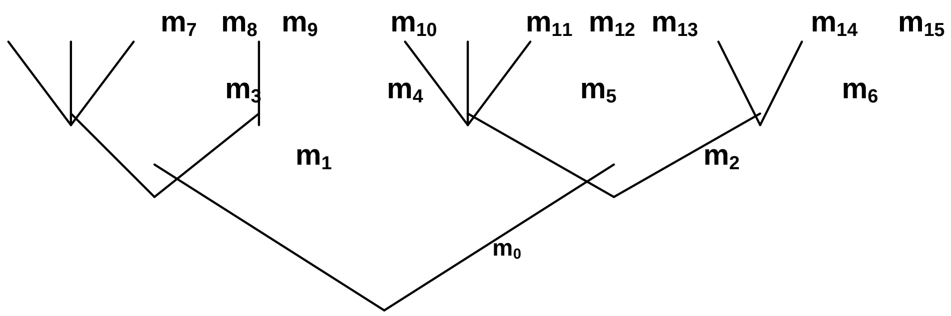
- As one might expect, two propositions *are identical* when they have the same structure of constituents and they are true in the same possible circumstances according to the same possible denotation assignment. Such a finer criterion of propositional identity explains why many strictly equivalent propositions have a different cognitive value. Propositions whose expression requires different acts of predication have a different structure of constituents. So necessarily true propositions about different objects (e.g the propositions that Cicero is Cicero and that Caesar is Caesar) are different. We do not have them in mind at the same moments. My criterion also distinguishes strictly equivalent propositions that we do not understand to be true in the same possible circumstances: these are not true according to the same possible denotation assignments to their constituents.

- My logic accounts for the fact that few necessarily true propositions are *pure tautologies* that are also *a priori* known to be true. In my approach, a proposition is *necessarily true* when it is true in every possible circumstance according to the real denotation assignment. In order to be *tautologically true*, a proposition has to be true in every possible circumstance according to all possible denotation assignments to its constituents. Unlike the proposition that whales are whales, the necessarily true proposition that whales are mammals is not a pure tautology. It is false according to possible denotation assignments

according to which whales are fishes. So we can believe that it is false. We now can distinguish in logic subjective and objective possibilities. By definition, a proposition is *subjectively possible* when it is true in a possible circumstance according to at least one possible denotation assignment. In order to be *objectively possible* a proposition has to be true in a possible circumstance according to the real denotation assignment.

Analysis of possible circumstances

In the logic of attitudes and action, the set of possible circumstances is provided with a ramified temporal structure. Human agents are free. Their attitudes and actions are not determined. When they do or think something, they could have done or thought something else. For that reason, one needs a ramified conception of time compatible with indeterminism. In branching time, a *moment* is a complete possible state of the actual world at a certain instant and the *temporal relation* of *anteriority / posteriority* between moments is partial rather than linear because of indeterminism. On the one hand, there is a single causal route to the past: each moment *m* is preceded by at most one chain of past moments. And all moments are historically connected: any two distinct moments have a common historical ancestor in their past. On the other hand, there are multiple future routes: several incompatible moments might follow upon a given moment. For facts, events or actions occurring at a moment can have incompatible future effects. Consequently, the set of moments of time has the formal structure of a *tree-like frame* of the following form:



A maximal chain *h* of moments of time is called a *history*. It represents a *possible course of history of our world*. When the history has a first and a last moment, the world has according to it a beginning and an end. A *possible circumstance* is a pair of a moment *m* and of a history *h* to which that moment belongs. Thanks to histories logic can analyze important modal notions like settled truth and historic necessity and possibility in addition to temporal notions. Certain propositions are true at a moment according to all histories. Their truth is then *settled at that*

moment[33] no matter how the world continues after it. So are past propositions because the past is unique. Their truth does not depend at all on histories. So are also propositions attributing attitudes to agents. Whoever believes or desires something at a moment then believes or desires that thing no matter what happens later. Contrary to the past, the future is open. The world can continue in various ways after indeterminist moments. Thus the truth of future propositions is not settled at each moment; it depends on which historical continuation *h* of that moment is under consideration. As Belnap [2001] pointed out, the future proposition that it will be the case that P (in symbols *Will*P) is true at a moment *m* according to a history *h* when the proposition that P is true at a moment *m′* posterior to *m* according to that very history. When there are different possible historic continuations of a moment, its actual future continuation is not then determined. However, as Occam pointed out, if the world continues after that moment, it will continue in a unique way. The actual historic continuation of each moment is then unique. Indeterminism does not prevent that unicity. Let h_m be the proper history of moment *m*. If *m* is the last moment of a history *h*, that history is then its proper history h_m. If on the contrary that moment continues, then all moments of its proper history have the same real historic continuation. Among all possible courses of history of this world, one will be its *actual course of history*. It is by hypothesis the proper history of the present actual moment *now*. From now on, I will say that a proposition is *true at a moment* according to a possible denotation assignment when it is then true in the history of that moment according to that assignment.

3. My new approach in the logic of attitudes

The notion of psychological mode is too rich to be taken as a primitive notion. Like Descartes, I consider that the two traditional categories of cognition and volition are essential components of psychological modes. But they divide into other components that we must take into consideration. In my view complex modes of attitudes are obtained from the primitive modes of belief and desire by adding special cognitive or volitive modes, special propositional or preparatory conditions. Here I am only concerned with the primitive attitudes of belief and desire. However, all cognitive modes contain beliefs and all volitive modes desires. So I want to formulate a very general theory of belief and of desire explicating traditional **categories of cognition and volition**. In particular, I advocate, like Searle, a very general explication of volition applying to all kinds of desires directed towards the past (shame), the present (lust) as well as the future (aspiration), even to desires known or believed to be satisfied (pleasure, joy) or unsatisfied (disappointment, regret), including desires directed at past actions that the agent would wish not to have done (remorse).

33 The terminology comes from Belnap [1992].

Beliefs and desires are directed to facts of the world and they have satisfaction conditions. In order that a belief or desire is satisfied, there must be a correspondence between the agent's ideas and things in the world. According to philosophy of mind, *beliefs* have the proper *mind-to-world direction of fit*. Whoever possesses a belief represents how things are in the world. A belief is satisfied when the agent's ideas corresponds to things as they are in the world. On the other hand, *desire* has the opposite *world-to-mind direction of fit*. Volitive attitudes are satisfied only if things in the world are or become as the agent desires them to be. Things in the world must then fit the agent's ideas. Each direction of fit between mind and the world determines which side is at fault in case of dissatisfaction. When a belief turns out to be false, it is the agent who is at fault, not the world. He should have had other ideas about the world. In such a case, the agent easily corrects the situation in changing his ideas. He adopts other beliefs. This is why the *truth predicates* (to be true and to be false) characterize so well *satisfaction* and *dissatisfaction* in the case of belief and other *cognitive* attitudes. However, such truth predicates do not apply to desire and other *volitive* attitudes whose direction of fit goes from things to words. For the world and not the agent is then at fault in the case of dissatisfaction of such attitudes. In the volitive case, the agent in general keeps his ideas and remains dissatisfied. Most often, agents having a *volitive attitude desire the fact represented by the propositional content no matter how that fact turns to be existent in the world*. For that reason, most volitive attitudes are *satisfied* when their propositional content is then true, no matter for which reason. Things are then such that the agent desires them to be, no matter what is the cause of their existence. In the present general explication of cognition and volition, desires like beliefs that agents possess at a moment are satisfied when their propositional content is true at that moment.

I will define the relation of compatibility with the truth of beliefs and the realization of desires of agents in a new way: First of all, in my approach, beliefs, desires and other attitudes of human agents are about objects that they represent under concepts. Each agent has in mind a certain set of attributes and concepts at each moment. (That set is empty when the agent does not exist or is totally unconscious at that moment.) Whoever believes or desires the existence of a fact has in mind all attributes and concepts of a propositional content that represents that fact. Otherwise, he would be unable to determine under which conditions his belief or desire is satisfied. As Wittgenstein and Searle pointed out, an attitude with entirely undetermined satisfaction conditions would be an attitude without real content; it would not be a real attitude. No one can believe or desire to be janissary without understanding the property in question.

Secondly, possible denotation assignments to propositional constituents rather than possible circumstances are compatible with the satisfaction of attitudes of agents. So to each agent *a* and moment *m* there corresponds in each model *a unique set Belief(a,m) of possible denotation assignments to attributes and concepts that are compatible*

with the truth of beliefs of that agent at that moment according to that model. By hypothesis, *Belief*(*a,m*) is the whole set *Val* of all possible denotation assignments to senses when the agent *a* has no sense at all in mind at the moment *m*. In that case, that agent has then no attitudes. Otherwise, *Belief*(*a,m*) is a *non empty proper* subset of *Val*. Whenever an agent has in mind propositional constituents, he always respects meaning postulates governing them and *there always are possible denotation assignments compatible with what that agent then believes*. Consider properties that an agent has in mind at a moment. When that agent has no idea at all about the denotation of a property in a circumstance then any possible denotation of a set of individuals under concepts is compatible with what he then believes. Suppose on the contrary that only such and such individuals under concepts could possess that property in that circumstance according to that agent. Then only possible denotation assignments associating with that property in that circumstance a set containing at least one of these individuals under concepts are compatible with what he then believes. And similarly for all other cases.

In my view, an agent *a believes a proposition P* at a moment *m* when he or she has then in mind all its constituent senses and that proposition is true at that moment according to all possible denotation assignments belonging to *Belief (a,m)*. We all ignore what will happen later in this world. But we now have a lot of beliefs directed at the future. As Occam pointed out, such beliefs are true when things will be as we believe in *the actual future continuation* of the present moment. Other possible historic continuations do not matter. The same holds for desire. In order that a present desire directed at the future is satisfied, it is not enough that things will be at a posterior moment as the agent now desires. They must be so later in the actual historic continuation of this world.

I will analyze desire according to the same approach. To each agent *a* and moment *m* there corresponds in each model a unique non empty set *Desire*(*a,m*) of possible denotation assignments to attributes and concepts that are compatible with the satisfaction of all desires of that agent at that moment. There is however an important difference between desire and believe. We can believe, but we cannot desire, facts without believing that they could not exist. For any desire contains a *preference*. Whoever desires something distinguishes two different ways in which represented things could be in the actual world. In a first preferred way, things are in the world as the agent desires them to be, in a second way, they are not. In the first case, the agent's desire is realized, in the second case, it is unrealized. Thus in order that an agent *a desires a proposition* at a moment *m*, it is not enough that he or she has then in mind all its constituents and that proposition is true at that moment according to all possible denotation assignments belonging to *Desire(a,m)*. That proposition must also not be tautological according to that agent. It must be false at a moment in a history according to possible denotation assignments.

My logic of belief and desire is compatible with philosophy of mind. Given new meaning postulates and its theory of truth, it accounts for the fact that *human agents are not perfectly rational*. We do not have in mind all concepts and attributes. So we ignore logical as well as necessary truths. Our knowledge is limited: we ignore the denotation of certain properties in many circumstances. In that case assignments associating different denotations to these properties in these circumstances are then compatible with our beliefs. We have false beliefs and unsatisfied desires. So many possible denotation assignments compatible with our beliefs and desires do not assign real denotations to attributes that we have in mind. Some of these assignments can even violate essential properties of our objects of reference. In that case we believe necessarily false propositions and desire impossible things. Each human being has a unique mother who gave birth to him or her. However that essential property does not correspond to any meaning postulate. Certain adopted children believe that their adoptive parents are their natural father and mother. One can easily account for such necessarily false beliefs in my analysis. Many traditional epistemic paradoxes are solved.

However *human agents are never totally irrational* according to my approach. On the contrary, they are *minimally rational* in a well determined way. First of all, they cannot believe or desire everything since some possible denotation assignments are always compatible with the satisfaction of their beliefs and desires. Moreover, they cannot possess certain beliefs and desires without possessing others. For all possible denotation assignments compatible with their beliefs and desires and consequently the contents of such attitudes have to respect meaning postulates. So human agents are in a sense *minimally omniscient* as regards logical truth; they cannot have in mind a pure tautology without knowing for certain that it is necessarily true. Represented things could not be in another way according to them. Similarly, so called *pure contradictions* (that is to say negations of tautologies) are false in every possible circumstance according to any possible denotation assignment. So we can neither believe nor desire contradictory things. Things could never be in a contradictory way according to us. Actual logicians still hope that arithmetic is complete (a necessarily false proposition if Gödel's proof is right). But no one could believe and desire both the completeness and the incompleteness of arithmetic (a pure contradiction). Sometimes we desire something (to be somewhere at a moment) for one reason and another incompatible thing (to be elsewhere at the same moment) for another reason. But when the logical form of such attitudes is fully analyzed, they are not desires with a contradictory content. Although agents believe all tautological propositions that they have in mind, they cannot desire the truth of such tautologies. As I said earlier, in order to desire a fact one must believe that it could not be the case. One can desire to drink; one can also desire not to drink. But one cannot desire to drink or not drink.

Incidentally, there is in predicative logic a new *strong* propositional *implication* much finer than Lewis' strict implication that is important for the analysis of psychological commitment. Let us say that a proposition P *strongly implies* another Q when firstly it contains the same or more predications than Q and it is tautologically true that if P then Q. Strong implication is finite, paraconsistent, decidable and *a priori known*.[34] As one might expect, any agent who believes a proposition also believes any proposition that is strongly implied by it. For that agent cannot have in mind the first proposition without having in mind the second and without understanding that the first cannot be true unless the second is. Unlike belief, desire is not closed under strong implication. Whoever desires to drink does not desire to drink or not drink.

I have just explained my way to analyze conditions of possession and satisfaction of beliefs and desires. In my view, the two traditional categories of cognition and volition correspond to the two families of compatibility relations $Belief^{a}_{m}$ and $Desire^{a}_{m}$ that exist between agents and moments, on one hand, and possible denotation assignments, on the other hand. We of course have real attitudes about objects at certain moments in this actual world. But we could have had other attitudes about the same or even about other objects. Often other agents attribute to us attitudes that we do not have. So compatibility relations $Belief^{a}_{m}$ and $Desire^{a}_{m}$ are moreover relative to possible denotation assignments in models. Agents' attitudes can differ according to different possible denotation assignments. First of all, each agent could have many concepts and attributes in mind. So in each model, every agent *a* has in mind at each moment *m* a *certain set val(a,m) of propositional constituents* according to any possible denotation assignment *val*. When this set is not empty, the agent then possesses beliefs and desires according to the assignment in question.

By hypothesis, $Belief^{a}_{m}(val)$ is the non empty set of all possible denotation assignment that are compatible with the truth of beliefs that agent *a* has at moment *m* according to denotation assignment *val*. Similarly, $Desire^{a}_{m}(val)$ is the non empty set of all possible denotation assignments that are compatible with the satisfaction of desires that agent *a* has at moment *m* according to assignment *val*. Of course, $Belief^{a}_{m}$ (*val*) and $Desire^{a}_{m}(val)$ are the whole set *Val* of all possible denotation assignment when there is no propositional constituent in the set *val(a,m)*. In my approach, an agent *a believes* or *desires* a proposition at a moment *m* (no matter what is the history) according to a possible denotation assignment *val* when firstly, he then has in mind all its concepts and attributes (they belong to the set *val*(*a,m*)) and secondly

34 See chapter 10 "Propositional Identity, Truth According to Predication and Strong Implication" in my book Logic, Thought and Action Springer 2005

that proposition is then true at that moment according to all possible assignments of the set $Belief^{a}_{m}(val)$ or $Desire^{a}_{m}(val)$. In the case of desire, the desired proposition must moreover be false in a circumstance according to the agent at the moment *m*.

As in modal logic, formal properties of psychological compatibility relations that correspond to attitudes depend on their nature. Whoever has a belief believes that he has that belief. So the relation $Belief^{a}_{m}(val)$ is transitive in each model. On the contrary, we often feel desires that we would wish not to feel. The relation $Desire^{a}_{m}$ is then not transitive. Some of our beliefs are false; many of our desires are unsatisfied. The compatibility relations $Belief^{a}_{m}(val)$ and $Desire^{a}_{m}(val)$ are also not reflexive. They are moreover not symmetric.

4. Formal semantics for belief and desire

The **vocabulary of the ideal object language** L of my elementary logic of attitudes contains a series of *individual constants* naming individual objects or agents, for each positive natural number n, a series of *predicates of degree* n expressing relations of degree n and the *syncategorematic expressions*: $\wedge$, $\neg$, *Settled*, $\Box$, *Actually*, *Tautological*, $\geq$, *Bel*, *Des* , *Was*, *Will*, (and).

Rules of formation of the ideography L

If R_n is a predicate of degree n and t_1, ..., and t_n are n individual terms, then $(R_n t_1 ... t_n)$ is a propositional formula. If a an individual term and A_p and B_p are propositional formulas, then $\neg A_p$, $\Box A_p$, *Tautological*A_p *Bel*aA_p , *Des*aA_p , $(A_p \geq B_p)$ and $(A_p \wedge B_p)$ and are new complex propositional formulas. $(R_n t_1 ... t_n)$ expresses the *elementary proposition* which predicates the attribute expressed by R_n of the n individuals under concepts expressed by t_1 , ..., t_n in that order. $\neg A_p$ expresses the *negation* of the proposition that A_p. *Settled*A_p expresses the *modal proposition* that it is settled that A_p. $\Box A_p$ expresses the *modal proposition* that it is then necessary that A_p. *Actually*A_p expresses the *proposition* that it is actually true that A_p.[35] *Was*A_p and *Will*aA_p respectively express the temporal *proposition* that it was and that it will be the case that A_p . *Bel*aA_p and *Des*aA_p respectively express the *proposition* that the individual named by a believes and desires that A_p . $(A_p \wedge B_p)$ expresses the *conjunction* of the two propositions that A_p and that B_p . $(A_p \supseteq B_p)$ means that all predications made in expressing the proposition that B_p are predications made in expressing the proposition that A_p. Finally, *Tautological*A_p means that the proposition expressed by A_p is tautological.

---

35 The proposition ActuallyAp is true at a moment according to a history when it is true at that moment in its proper history.

Rules of abbreviation

I will use standard rules of abbreviation for elimination of parentheses and truth or modal connectives of *disjunction* $\vee$, *material implication* $\Rightarrow$, *material equivalence* $\Leftrightarrow$, *possibility* and *strict implication* $—\!\in$. Here are new rules:

Same structure of constituents: $A_p \equiv B_p =_{df} (A_p \geq B_p) \wedge (B_p \geq A_p))$

Identical individual concepts: $\hat{}t_1 = \hat{}t_2 =_{df} (R_1\, t_1) \geq (R_1 t_2)$ where R_1 is the first unary predicate

Identical attributes: $\hat{}R_n = \hat{}R'_n =_{df} (R_n\, t_1...t_n) \geq (R'_n\, t_1...t_n)$

*Always*A $=_{def}$ $\neg$*Was*$\neg$A $\wedge$ A $\wedge$ $\neg$*Will*$\neg$A

Historical possibility: $\Diamond$A $=_{def}$ $\neg\Box\neg$A

Universal Necessity $\Box$A $=_{def}$ *Always*$\Box$A $\wedge$ $\Box$*Always*A

Analytic implication[36]: $(A_p \rightarrow B_p) =_{def} (A_p \geq B_p) \wedge (A_p —\!\in B_p)$

Strong implication: $A_p \mapsto B_p =_{df} (A_p \geq B_p) \wedge$ *Tautological* $(A_p \Rightarrow B_p)$

Propositional identity: $A_p = B_p =_{df} (A_p \mapsto B_p) \wedge B_p \mapsto A_p$

Strong psychological commitment *Bel*A_p ▶ *Bel*B_p $=_{df}$ *Tautological*(*Bela*A_p $\Rightarrow$ *Bela*B_p)

Weak psychological commitment *Bel*A_p $\rhd$ *Bel*B_p $=_{df}$ *Tautological*(*Bela*A_p $\Rightarrow \neg$*Bela*$\neg B_p$)

And similarly for *Des*.

Definition of a model structure

A *standard model* $\mathfrak{M}$ for **L** is a structure < *Moments*, *Individuals, Agents*, *Concepts*, *Attributes, Val*, *Predications*, *Belief*, *Desire*, * , $\otimes$, $\equiv$>
that satisfies the following conditions:

- The set *Moments* is a set of *moments of time*. It is partially ordered by a temporal relation $\leq$ as in ramified temporal logic. $m_1 < m_2$ means that moment m_1 is *anterior to* moment m_2. By definition, < is subject to *historical connection* and *no downward branching*. Any two distinct moments have a common historical ancestor. Moreover, the past is unique. A maximal chain *h* of moments is called a *history*. It represents a possible course of history of the world. The set *Circumstances* of all *possible circumstances* contains all pairs *m/h* where *m* is a moment belonging to the history *h*. Among all histories to which belongs a moment *m* there is one h_m representing how the world would continue after that moment. If $m' \in h_m$, $h_{m'} = h_m$. The set *Instants*, whose elements ι, ι',... are called *instants*, is a partition of the set *Time* which satisfies *unique intersection* and *order preservation*. So to any instant ι and history *h* there corresponds a unique moment $m(\iota,h)$ belonging to both ι and *h*. And $m(\iota,h) \leq m(\iota',h)$ when $m(\iota,h') \leq m(\iota',h')$. Two moments of time *m* and

36 See W.T. Parry, "Ein Axiomsystem fuer eine neue Art von Implikation (analytische Implication)", Ergebnisse eines Mathematisches Colloquiums, Volume 4, 1933 and "Comparison of entailment theories", The Journal of Symbolic Logic, Volume 37, 1972 as well as K. Fine, "Analytic implication", in Notre Dame Journal of formal Logic, Volume 27, number 2, April 1986.

m' are *coinstantaneous* (in symbols: $m \cong m'$) when they belong to the same instant. Coinstantaneous moments *m* and *m' represent* two complete possible states of the world in which things could be at a certain instant.

- *Individuals* is a set of possible *individual objects*. For each moment *m*, $Individuals_m$ is the set of individuals existing at that moment. *Agents* is a subset of *Individuals* containing *persons.*

- *Concepts* is the set of *individual concepts* and *Attributes* is the set of *attributes* of individuals considered in model □. For each natural number n, *Attributes*(n) is the subset of *Attributes* containing all *attributes* of degree n .

- The set *Val* is a proper subset of $((Concepts \times Circumstances) \rightarrow (Individuals \cup \{\emptyset\})) \cup \bigcup_n ((Attributes(n) \times Circumstances) \rightarrow$ □$(Concepts^n))$. *Val* contains all possible *denotation assignments* of the model M. Such assignments are also called possible *valuations of constituents*. For any possible circumstance *m/h*, $val\,(c_e, m/h) \in$ *Individuals* when individual concept c_e has a denotation in the circumstance *m/h* according to assignment *val*. Otherwise $val\,(c_e, m/h) = \emptyset$. For any attribute R_n of degree n, $val\,(R_n, m/h) \in$ □$(Concepts^n)$. The set *Val* contains a *real valuation* val□ which assigns to concepts and attributes their *actual denotation* in each possible circumstance according to the model □. Moreover, there corresponds to each agent *a*, moment *m* and assignment *val* a particular set *val*(*a*,*m*) containing all propositional constituents that the agent *a* has in mind at that moment according to that assignment.

- *Belief* is a function from *Agents* × *Moments* × *Val* into □ (*Val*) that associates with any agent *a*, moment *m* and valuation val, a non-empty set $Belief^a_m(val) \subseteq Val$ containing all possible denotation assignments which are compatible with the beliefs that agent *a* has at the moment *m* according to that valuation. The relation of epistemic compatibility corresponding to $Belief^a_m$ is transitive. Moreover, $val(a,m) \subseteq val'(a,m)$ when $val' \in Belief^a_m(val)$.

- *Desire* is a function from *Agents* × *Moments* × *Val* into □ (*Val*) that associates with any agent *a*, moment *m* and valuation val, a non empty set $Desire^a_m(val) \subseteq Val$ containing all possible denotation assignments which are compatible with the desires that agent *a* has at moment *m* according to that valuation.

As one can expect, $Val = Belief^a_m(val) = Desire^a_m(val)$ when $val(a,m) = \emptyset$.

- The set *Predications* is a subset of □(*Attributes* ∪ *Concepts*) containing all sets of propositional constituents with which *predications* that can be made in the language **L**. Each member of that set is a set of the form $\{R_n, c^1_e, \ldots, c^n_e\}$ containing a single attribute R_n of degree n and a number k of individual concepts $k \leq n$ with $k \neq 0$ when n is positive. The power set □*Predications* is closed under union ∪ , a modal unary operation *

and, for each agent concept a, a unary epistemic operation $\otimes_a$ of the following form: For any Γ, Γ_1 and $\Gamma_2 \in$ □*Predications*, $\Gamma \subseteq {}^*\Gamma$ and ${}^*\Gamma \subseteq \otimes_a \Gamma$. Moreover, ${}^*(\Gamma_1 \cup \Gamma_2) = {}^*\Gamma_1 \cup {}^*\Gamma_2$ and ${}^{**}\Gamma = {}^*\Gamma$. Similarly, $\otimes_a (\Gamma_1 \cup \Gamma_2) = \otimes_a \Gamma_1 \cup \otimes_a \Gamma_2$ and $\otimes_a \otimes_a \Gamma = \otimes_a \Gamma$. By definition, when $\cup\Gamma \subseteq val(a,m)$, $\otimes_a \Gamma \subseteq val(a,m)$.[37]

- Finally, $\equiv\ \equiv$ is a function that associates with each expression of □ the sense of that expression in the possible interpretation □. $\equiv A \equiv$ satisfies the following clauses:
- For any individual constant of L $\equiv c_\alpha \equiv\ \in$ *Concepts*.
- For any predicate R_n of degree n, $\equiv R_n \equiv\ \in$ *Attributes*(n).
- For any propositional formula A_p, $\equiv A_p \equiv$ belongs to a subset of □*Predications* $\times$ (*Circumstances* $\rightarrow$ □*Val*). Remember that each proposition P has two essential features: the set id_1P of all its predications and the set id_2P of all possible denotation assignments according to which it is true. Consequently:
- $id_1\equiv[(R_n c_1 ,...,c_n)]\equiv\ = \{\{\equiv R_n\equiv, \equiv t_1\equiv,...,\equiv t_n\equiv\}\}$ and $id_2\equiv[(R_n c_1 ,...,c_n)]\equiv(m/h) = \{ f \in Val \,/\, <\equiv c_1\equiv,\ldots,\equiv c_n\equiv)) \in f(\equiv R_n\equiv, (m/h))\}$.
- $id_1\equiv\neg A_p\equiv\ = id_1\equiv A_p\equiv$ and $id_2\equiv\neg A_p\equiv(m/h) = Val - id_2\equiv A_p\equiv(m/h)$.
- $id_1\equiv\Box A_p\equiv\ = {}^*\, id_1\equiv A_p\equiv$ and $id_2\equiv\Box A_p\equiv(m/h) = \bigcap_{m'\equiv m} \bigcap_{m'\in h'} id_2\equiv A_p\equiv(m'/h')$.
- $id_1\equiv Will A_p\equiv\ = {}^*\, id_1\equiv A_p\equiv$ and $id_2\equiv Will A_p\equiv(m/h) = \bigcup_{m'>m} id_2\equiv A_p\equiv(m'/h)$.
- $id_1\equiv Was A_p\equiv\ = {}^*\, id_1\equiv A_p\equiv$ and $id_2\equiv Was A_p\equiv(m/h) = \bigcup_{m'<m} id_2\equiv A_p\equiv(m'/h)$.
- $id_1\equiv Settled A_p\equiv\ = {}^*\, id_1\equiv A_p\equiv$ and $id_2\equiv Settled A_p\equiv(m/h) = \bigcap_{m\in h} id_2\equiv A_p\equiv(m/h)$.
- $id_1\equiv Actually A_p\equiv\ = {}^*\, id_1\equiv A_p\equiv$ and $id_2\equiv Actually A_p\equiv(m/h) = id_2\equiv A_p\equiv(m/h_m)$.
- $id_1\equiv Tautological A_p\equiv\ = {}^* id_1\equiv A_p\equiv$ and $id_2\equiv Tautological A_p\equiv(m/h) = \bigcap_{m'/h'} id_2\equiv A_p\equiv (m'/h')$.
- $id_1(\equiv B_p \wedge C_p\equiv) = id_1(\equiv B_p\equiv) \cup id_1(\equiv C_p \equiv)$; $id_2\equiv B_p \wedge C_p\equiv(m/h) = id_2\equiv B_p\equiv(m/h) \cap id_2\equiv C_p \equiv(m/h)$.
- $id_1(\equiv B_p \geq C_p\equiv) = {}^*(id_1(\equiv B_p\equiv) \cup id_1(\equiv C_p\equiv))$ and $id_2\equiv B_p \geq C_p\equiv(m/h) = Val$ when $id_1\equiv B_p\equiv\ \subseteq id_1\equiv C_p\equiv$. Otherwise, $id_2\equiv B_p \supseteq C_p \equiv(m/h) = \emptyset$.
- $id_1\equiv Bel a B_p\equiv\ = \otimes_{\Pi a \Pi} id_1\equiv B_p\equiv$ and $id_2\equiv Bel a B_p\equiv(m/h) = \{val \,/\, \cup\equiv B_p\equiv\ \subseteq val(a,m)$ and $((\equiv A_\mu\equiv(val) \cap Belief^{a}_{m_c}(val)) \subseteq id_2\equiv B_p\equiv(m/h_m)$ where $a = val\equiv a\equiv(m,h)\}$. And similarly for $\equiv Des a B_p\equiv^{\sigma}_{c}$ with the additional condition that for some m and h, $id_2\equiv B_p\equiv(m/h) \neq Val$.

37 As one can expect, each agent who has attitudes has attitudes about himself.

Definition of truth and validity

A propositional formula A_p of **L** is *true* in a possible circumstance *m/h* according to a standard model □ if and only $\equiv A_p \equiv$ is true in *m/h* according to *val*□. The formula A_p is *valid* (in symbols: $\models A_p$) when it is true in all possible circumstances according to all standard models.

4. An axiomatic system

I conjecture that one can prove all and only valid formula of L containing syncategorematic symbols $\wedge$, $\neg$, □ , *Tautological*, $\geq$, *Bel*, *Des*, (and) in the following axiomatic system □:

Axioms

The axioms of □ are all the instances in that sub-language of L of classical axiom schemas of truth functional logic and S5 modal logic as well as instances of the following new schemas:

Axiom schemas for tautologies

(T1) $\textit{Tautological}A_p \Rightarrow \Box A_p$
(T2) $\textit{Tautological}\ A_p \Rightarrow \textit{Tautological}\ \textit{Tautological}A_p$
(T3) $\neg\textit{Tautological}\ A_p \Rightarrow \textit{Tautological}\neg\textit{Tautological}A_p$
(T4) $\textit{Tautological}A_p \Rightarrow (\textit{Tautological}\ (A_p \Rightarrow B_p) \Rightarrow \textit{Tautological}B_p)$
(T5) $(A_p \geq B_p) \Rightarrow \textit{Tautological}(A_p \geq B_p)$
(T6) $\neg(A_p \geq B_p) \Rightarrow \textit{Tautological}\neg(A_p \geq B_p)$

Axiom schemas for propositional identity

(I1) $A_p = A_p$
(I2) $(A_p = B_p) \Rightarrow (C \Rightarrow C^*)$ where C^* and C are propositional formulas which differ at most by the fact that an occurrence of B_p in C^* replaces an occurrence of A_p in C.
(I3) $(A_p = B_p) \Rightarrow \textit{Tautological}\ (A_p = B_p)$
(I4) $\neg(A_p = B_p) \Rightarrow \textit{Tautological}\ \neg(A_p = B_p)$

Axiom schemas for belief

(B1) $(\textit{Bela}A_p \wedge \textit{Bela}B_p) \Rightarrow \textit{Bela}(A_p \wedge B_p)$
(B2) $\textit{Tautological}A_p \Rightarrow \neg\textit{Bela}\neg A_p$
(B3) $\textit{Tautological}A_p \Rightarrow (\textit{Bela}A_p \Rightarrow \textit{BelaTautological}A_p)$
(B4) $\textit{Bela}A_p \Rightarrow ((A_p \mapsto B_p) \Rightarrow (\textit{Bela}B_p))$
(B5) $\textit{Bela}A_p \Leftrightarrow (\textit{BelaBela}A_p)$
(B6) $\textit{Bela}A_p \Rightarrow \textit{Bela}\Diamond A_p$

Axiom schemas for desire

(D1) $(\textit{Desa}A_p \wedge \textit{Desa}B_p) \Rightarrow \textit{Desa}(A_p \wedge B_p)$
(D2) $\textit{Tautological}A_p \Rightarrow \neg(\textit{Desa}A_p \vee \textit{Desa}\neg A_p)$
(D3) $\textit{Desa}A_p \Rightarrow (((A_p \mapsto B_p) \wedge \neg\textit{Tautological}A_p)) \Rightarrow (\textit{Desa}B_p)))$

(D4) $\mathit{Desa}A_p \Rightarrow \mathit{Bela}\neg \mathit{Tautological}A_p$

(D5) $\mathit{Desa}A_p \Rightarrow \mathit{Desa}\Diamond A_p \wedge \mathit{Bela}\neg\Box A_p)$

Axiom schemas for propositional composition

(C0) $((R_n t_1,\ldots,t_n) > A_p) \Rightarrow (A_p = (R_n t_1,\ldots,t_n))$
(C1) $A_p \geq A_p$
(C2) $(A_p \geq B_p) \Rightarrow ((B_p \geq C_p) \Rightarrow (A_p \geq C_p))$
(C3) $(A_p \wedge B_p) \geq A_p$
(C4) $(A_p \wedge B_p) \geq B_p$
(C5) $((C_p \geq A_p) \wedge (C_p \geq B_p)) \Rightarrow C_p \geq (A_p \wedge B_p)$
(C6) $A_p \equiv \neg A_p$
(C7) $\Box A_p \equiv \mathit{Tautological}A_p$
(C8) $(A_p \geq B_p) \equiv \Box(A_p \wedge B_p)$
(C9) $\Box A_p \geq A_p$
(C10) $\mathit{Bela}A_p \geq \Box A_p$ And similarly for *Des*a.
(C11) $\Box\neg A_p \equiv \Box A_p$ And similarly for *Bel*a and *Des*a. (C14)
(C12) $\Box(A_p \wedge B_p) \equiv (\Box A_p \wedge \Box B_p)$ And similarly for *Bel*a and *Des*a. (C16)
(C13) $\Box\Box A_p \equiv \Box A_p$ And similarly for *Bel*a and *Des*a. C18)

Rules of inference

The two rules of inference of my axiomatic system are:
The *rule of Modus Ponens*: (MP) From sentences of the form A and (A $\Rightarrow$ B) infer B .
The *tautologization rule*: (RT) From a theorem A infer *Tautological*A .

5. Valid laws

Here are important valid laws of my logic of belief and desire.[38]

Laws of structure of constituents A proposition has all elementary propositions of its arguments. $\models A_p \geq (R_n t_1,\ldots,c_n)$ when $(R_n t_1,\ldots,c_n)$ occurs in A_p. Modal and epistemic propositions have in general more elementary propositions than their arguments. Thus $\Box A_p \geq \Box A_p$ and $\Box \Box A_p \geq \mathit{Bela}A_p$.

Laws for tautologyhood Tautologyhood is stronger than necessary truth and contradiction stronger than necessary falsehood. $\models (\mathit{Tautological}A_p) \Rightarrow \Box A_p$. But $\Box \Box A_p \Rightarrow \mathit{Tautological}A_p$ Some tautologies are modal and epistemic. Thus $\models \mathit{Tautological}(\Box A_p \Rightarrow A_p)$.

Agents are minimally rather than perfectly rational.

They do not believe all necessary truths and they can believe and desire necessarily false propositions. Thus $\Box \Box A_p \Rightarrow \mathit{Bela}A_p$ and $\Box \neg\blacklozenge A_p \Rightarrow$

38 Some of these laws are stated in my paper "Truth, Belief and Certainty in Epistemic Logic" in E. Maier et al Proceedings of Sinn und Bedeutung 9 NCS Nijmegen 2005

*Bel*a $\neg A_p$. However they are minimally consistent: they cannot believe that a tautology is false or that a contradiction is true. $\models$*Tautological*$\neg A_p \Rightarrow \neg$*Bela*$A_p$ Moreover they neither desire tautologies nor contradictions (axiom D3). Now in order to believe or desire a proposition an agent must have in mind its attributes and concepts. Unlike God, human agents do not have in mind all propositional constituents. Consequently they do not know or even believe all tautologies. $\Box$*Tautological*$A_p \Rightarrow$ *Bela*A_p. The limits of their language impose limits to their thoughts. However whenever they express a tautology and a contradiction, they know just by apprehending their logical form that the first is necessarily true and the second necessarily false (axiom B3).

Laws for tautological implication

Tautological implication is much finer than strict implication. $\models$ *Tautological* $(A_p \Rightarrow B_p) \Rightarrow (A_p \mathrel{—\!\!\in} B_p)$ But $\Box$ $(A_p \mathrel{—\!\!\in} B_p) \Rightarrow$ *Tautological* $(A_p \Rightarrow B_p)$. Necessarily true propositions are strictly implied by others. $\models$ $\Box A_p \Rightarrow (B_p \mathrel{—\!\!\in} A_p)$. But only tautologies can tautologically imply other tautologies. $\models$ ((*Tautological*B_p) $\wedge$ *Tautological* $(A_p \Rightarrow B_p)) \Rightarrow$ *Tautological* A_p. So $\Box$ $\Box A_p \Rightarrow$ *Tautological*$(B_p \Rightarrow A_p)$. Similarly, necessarily false propositions strictly imply all other propositions. $\models \Box\neg A_p \Rightarrow (A_p \mathrel{—\!\!\in} B_p)$. But only contradictions can tautologically imply contradictions. So $\Box$ $\Box\neg A_p \Rightarrow$ *Tautological*$(A_p \Rightarrow B_p)$.
Beliefs are not closed under tautological implication. $\Box$ (*Tautological* $(A_p \Rightarrow B_p)) \Rightarrow$ (*Bela*$A_p \Rightarrow$ *Bela*B_p)) Because $\Box$ (*Tautological* $(A_p \Rightarrow B_p)) \Rightarrow (A_p \geq B_p)$). However whoever believes a proposition cannot believe the negation of a proposition that the first tautologically implies. For the conjunction of both is a contradiction. This is why tautological implication generates *weak psychological and illocutionary commitment*. Any assertion that P *weakly commits* the agent to asserting any proposition Q that P tautologically imply according to illocutionary logic.[39] Similarly, $\models$ *Tautological* $(A_p \Rightarrow B_p) \Rightarrow$ (*Bel*$A_p \mathrel{—\!\!\in} \neg$*Bela*$\neg B_p$) and $\models$ *Tautological* $(A_p \Rightarrow B_p) \Rightarrow$ (*Desa*$A_p \mathrel{—\!\!\in} \neg$*Desa*$\neg B_p$)

Laws for strong implication

Strong implication is a stronger kind of implication than strict, tautological and analytic implications. It requires the same or a richer structure of constituents in addition to tautological implication. There are two reasons why a proposition can fail to strongly imply another. Firstly, that other proposition requires new predications. $\models \neg(A_p \geq B_p) \Rightarrow \neg(A_p \mapsto B_p)$. In that case, one can think the first proposition without thinking the second. Secondly, the first proposition does not tautologically imply the other. In that case one can ignore that it implies the other.

---

39 See "Success, Satisfaction and Truth in the Logic of Speech Acts and Formal Semantics" [2004]

So strong implication is finer than analytic implication. $\square$ $(A_p \rightarrow B_p) \Rightarrow (A_p \mapsto B_p)$ In particular $\square$ $(A_p \rightarrow B_p) \Rightarrow$ *Bela*$A_p \Rightarrow$ *Bela* B_p. Unlike strict and tautological implications, strong implication is anti-symmetric. Consequently, $\models A_p \mapsto B_p \Leftrightarrow ((A_p \wedge B_p) = A_p)$

Strong implication is *decidable*. For $\models A_p \geq B_p$ when all propositional constants which occur in B_p also occur in A_p. And $\models$ *Tautological* $(A_p \Rightarrow B_p)$ when any semantic tableau of S5 modal logic for $(A_p \Rightarrow B_p)$ closes.

Moreover, strong implication is *finite*: every proposition only strongly implies a finite number of others. For it contains a finite number of elementary propositions. In particular, a proposition only strongly implies tautologies having its elementary propositions. $\models$ *Tautological* $B_p \Rightarrow (A_p \mapsto B_p \Leftrightarrow A_p \geq B_p)$. A contradiction only strongly propositions having its elementary propositions. $\models$ *Tautological* $\neg A_p \Rightarrow (A_p \mapsto B_p \Leftrightarrow A_p \geq B_p)$

For all these reasons, strong implication is *known a priori*. $\models (A_p \mapsto B_p) \Rightarrow$ (*Bela*$A_p \Rightarrow$ *Bela*$(A_p \Rightarrow B_p)$). However $\mapsto$ does not obey the rule of *Modus Tollens*. $\square$ $(A_p \mapsto B_p) \Rightarrow (\neg B_p \mapsto \neg A_p)$. For $\square$ $(A_p \mapsto B_p) \Rightarrow (B_p \geq A_p)$. So $\square$ $(A_p \mapsto B_p) \Rightarrow$ (*Bela*$\neg B_p \Rightarrow$ *Bela*$\neg A_p$)

Natural deduction

Valid laws of inference of natural deduction generate strong implication when their premises contain all propositional constants of their conclusion. Here are some laws:

The law of introduction of belief: $\models A_p \mapsto B_p \Rightarrow$ *Bela*$A_p \mapsto$ *Bela*B_p

The law of introduction of desire: $\models ((A_p \mapsto B_p) \wedge \neg$*Tautological*$A_p) \Rightarrow$ *Desa*$A_p \mapsto$ *Desa*B_p

The law of elimination of conjunction: $\models (A_p \wedge B_p) \mapsto A_p$ and $\models (A_p \wedge B_p) \mapsto B_p$

The law of elimination of disjunction: $\models ((A_p \mapsto C_p) \wedge (B_p \mapsto C_p)) \Rightarrow (A_p \vee B_p) \mapsto C_p$

Failure of the law of introduction of disjunction: $\square$ $A_p \mapsto (A_p \vee B_p)$.

Consequently, strong implication is stronger than *entailment* which obeys the law of introduction of disjunction. Clearly $\square$ *Bela*$A_p \mapsto$ *Bela* $(A_p \vee B_p)$. Similarly, $\square$ *Desa*$A_p \mapsto$ *Desa*$(A_p \vee B_p)$.

The *law of introduction of negation*: $\models A_p \mapsto O_t \Rightarrow (A_p \mapsto \neg A_p)$ where O_t is any contradiction.

Failure of the law of elimination of negation: $\square$ $(A_p \wedge \neg A_p) \mapsto B_p$

Agents can have relatively inconsistent beliefs and desires. $\square$ $(A_p \text{—}\!\in \neg B_p) \Rightarrow \neg\blacklozenge$ *Bela* $(A_p \wedge B_p)$ Similarly, $\square$ $(A_p \text{—}\!\in \neg B_p) \Rightarrow \neg\blacklozenge$ *Desa* $(A_p \wedge B_p)$

They are *paraconsistent*. $\square$ $(A_p \text{—}\!\in \neg B_p) \Rightarrow$ (*Bela* $(A_p \wedge B_p) \Rightarrow$ *Bela* C_p)

But they always respect the principle of non contradiction. $\models \neg\blacklozenge$ *Bela*$(A_p \wedge \neg A_p)$

The *law of elimination of material implication*: $\models (A_p \wedge (A_p \Rightarrow B_p)) \mapsto B_p$

The *law of elimination of necessity*: $\models \square A_p \mapsto A_p$

The *law of elimination of possibility*: $\models \blacklozenge A_p \mapsto B_p \Rightarrow A_p \mapsto B_p$

Laws of propositional identity

All the classical *Boolean laws of idempotence, commutativity, associativity* and *distributivity* are valid laws of propositional identity: So $\models$ *Bela*A_p = *Bela*($A_p \wedge A_p$); $\models$*Bela* ($A_p \wedge B_p$) = *Bela*($B_p \wedge A_p$); $\models$*Bela*$\neg$($A_p \vee B_p$) = *Bela* ($\neg A_p \wedge \neg B_p$); $\models$*Bela* ($A_p \wedge$ ($B_p \vee C_p$)) = *Bela*(($A_p \wedge B_p$) $\vee$ ($A_p \wedge C_p$)) and $\models$*Bela* □($A_p \wedge B_p$) = *Bela* (□$A_p \wedge$ □B_p).

The classical laws of *reduction* are also valid: $\models \neg\neg A_p = A_p$ and $\models$ *BelaBela*A_p = *Bela*A_p However, □ *Desa*$A_p \Rightarrow$ *DesaDesa*A_p . Unlike hyperintensional logic, predicative logic does not require that identical propositions be *intensionally isomorphic*.[40] First of all, as I said earlier, the order of predication does not always affect truth conditions. Similarly, the order and number of applications of propositional operations does not always affect the logical form. Clearly, $\models$*Bela*($A_p \Leftrightarrow B_p$) = *Bela*($B_p \Leftrightarrow A_p$) Intensional isomorphism is too strong a criterion of propositional identity.

However, propositional identity requires more than the *co-entailment* advocated in the logic of relevance. As M. Dunn points out, it is unfortunate that A_p and ($A_p \wedge$ ($A_p \vee B_p$) co-entail each other.[41] For most formulas of such forms are not synonymous. Co-entailment is not sufficient for synonymy because it allows for the introduction of new sense. □ $A_p \mapsto$ ($A_p \wedge$ ($A_p \vee B_p$). Consequently □ *Bela* $A_p \mapsto$ *Bela* ($A_p \wedge$ ($A_p \vee B_p$).[42]

References

Anderson A. R., N. Belnap & J. M. Dunn: ***Entailment The Logic of Relevance and Necessity***, Princeton University 1992

Belnap N. & M. Perloff: "The Way of the Agent" in ***Studia Logica***, 51, 463-484, 1992

Belnap N., M. Perloff & Ming Xu: ***Facing the Future*** *Agents and Choices in Our Indeterminist World* Oxford University Press 2001

Carnap R.: ***Meaning and Necessity***, University of Chicago Press 1956

Cresswell M. J.: "Hyperintensional Logic", ***Studia Logica***, 34,: 25-38, 1975

R. Descartes ***Œuvres complètes*** La Pléiade Gallimard 1953

Frege G.: 1970, "On Sense and Reference" in P. Geach and M. Black (eds), ***Translations from the philosophical Writings of Gottlob Frege***, Blackwell

Hintikka J.: "Semantics for Propositional attitudes" in L. Linsky (ed) ***Reference and Modality*** Oxford University Press 1962

Barcan Marcus R.: ***Modalities*** , Oxford University Press, 1993

Lewis C. I.: ***A Survey of Symbolic Logic***, University of California Press, 1918

Parry W.T.: "Comparison of Entailment Theories", ***Journal of Symbolic Logic***, 37, 1972

A. Plantinga ***The Nature of Necessity***, Oxford University Press 1974

---

40 See Max J. Cresswell, "Hyperintensional Logic". Studia Logica [1975].

41 See his philosophical rumifications in Anderson et al [1992].

42 A general predicative logic of propositions dealing with generalization, logical and historic modalities, ramified time, attitudes and action is fully developed in my next book Logic Truth & Thought.

Vanderveken D.: ***Meaning and Speech Acts*****,** 1990, Volume 1: ***Principles of Language Use*** and 1991, Volume 2: ***Formal Semantics of Success and Satisfaction***, Cambridge University Press
– "Success, Satisfaction and Truth in the Logic of Speech Acts and Formal Semantics" in S. Davis & B. Gillan (eds) ***Semantics A Reader***, Oxford University Press, 710-734, 2004
– "Propositional Identity Truth According to Predication and Strong Implication With a Predicative Formulation of Modal Logic" in D. Vanderveken (ed) ***Logic, Thought & Action***, Springer,185-216, 2005
- "Truth, Belief and Certainty in Epistemic Logic" in E. Maier *et al* ***Proceedings of Sinn und Bedeutung 9*** NCS Nijmegen 2005
– Forthcoming, "Fondements de la logique des attitudes" in the special issue *Language and Thought* of ***Manuscrito***
- Forthcoming, ***Propositions, Truth & Thought***

INDUCTION WITH MACHINE LEARNING REVISITED

JEAN-GABRIEL GANASCIA
LIP6 – Laboratoire d'Informatique de Paris 6 (CNRS)
University Pierre and Marie Curie (Paris VI)
+33 (0) 1 44 27 37 27
Jean-Gabriel.Ganascia@lip6.fr

ABSTRACT

Machine Learning is mainly concerned with the construction of knowledge from experiences. In the past, different techniques have been developed to derive knowledge and/or to improve machines behaviour from flows or sets of examples. Since the aim of Machine Learning is to build general statements from facts (i.e. from particulars), the corresponding inference has been viewed as an induction.

However, it appears that neither induction viewed as an inversion of deduction, nor formal theories of induction can account for the structural induction as it has been simulated by Machine Learning techniques.

This paper explains why these theories of induction are not satisfactory. Then it shows that, coming back to Aristotle writings, we have to distinguish two approaches of induction: a logical approach, which tries to define the theoretical status of inductive reasoning and an empirical methodology developed for natural science which is more closely related to Machine Learning structural induction.

KEYWORDS

Machine Learning, Induction, Structural Induction, Formal Theory of Induction, Inductive Logic Programming (ILP), Inductive Logic

Machine Learning

Machine Learning and Induction

Machine Learning enables machines to learn all kinds of knowledge and know-how without any help. Since the beginning, at the time of the first cybernetic, in the forties, or later, in 1950, when the famous Turing's paper [Turing 50] about thinking machines was published, Machine

Learning has been viewed as a key issue in artificial intelligence. Up to now, for almost all artificial intelligence researchers, it always appeared that to be intelligent, machines need to acquire by themselves huge amount of knowledge about the outside world. As a consequence, during more than sixty years now, many machine learning algorithms have been designed, implemented and tested.
Whatever the learning mechanisms used in Machine Learning, be they similarity-based, generalization-based, explanation-based, action-based, reinforcement-based, instance-based, case- or analogy-based, neural networks or adaptive algorithms, etc., the aim is always to find ways for machines to learn from certain singular experiences and either to apply the results to other particular cases or to build knowledge, laws and theories. In all events, the reasoning starts from the observation of certain particular cases and derives either general knowledge that is interesting in itself, or problem-solving procedures to be used on similar cases to those used for the learning process. Since the inferences simulated by machine learning mechanisms move from the particular to the general, or to another particular, they correspond to what has been identified since ancient times by logicians and philosophers as "induction".
Moreover, many of the learning mechanisms used in artificial intelligence mimic inductive procedures identified by philosophers. It is the case for the detection of association rules, for the construction of decision trees, for the generalization of depiction by the use of the so-called « dropping rule », for the neural networks, for the Support Vector Machines and for most of the classical techniques developed in artificial intelligence.

Formal Theories of Learning

During the sixties, with E. Gold [Gold 67] “identification to the limit” learning paradigm, and more recently with the “Probably Approximately Correct” learning – PAC learning – paradigm [Valiant 84] or with the statistical learning theory [Vapnik 95], many researchers attempted to define the theoretical limitations of learning machines. Those formal theories aim at clarifying the mathematical characteristics of learning algorithms in computational terms, i.e. in terms of inputs, outputs and spatial versus temporal algorithmic complexity. In a way, they may be viewed as theories of inductive inference complementary to the inductive logics developed by Carnap [Carnap 62], Hintika and others. Within those inductive inference theories, induction is viewed as an approximative inference of which uncertainty has to continuously decrease with the number of observations. The logical status of induction is not really determined, but the mathematical properties – i.e. the number of required examples, the number of features, the speed of convergence etc. – of inductive inferences are well approached.

However, among the different machine learning mechanisms used in artificial intelligence to infer knowledge from facts, one of them seems to have no counterpart in the traditional approaches to induction; it is the matching mechanism which identifies those subparts of objects corresponding to each other. If objects are structurally described as being made up of subparts, and not only as sets of attributes, matching is required by all mechanisms that attempt to simulate induction. Since, up till now, depiction has been referred to in philosophical studies of induction only as sets of attributes, matching has never needed to be mentioned.

It is also the case with many other traditional approaches drawn either from data analysis, such as decision tree construction; or from cybernetics, such as neural networks; or from artificial life, such as genetic algorithms. It is only with the emphasis laid by symbolic artificial intelligence on knowledge representation that the role of structured descriptions has become central. Much research has been done on the mapping of subparts of examples. See, for instance, the work of Plotkin (Cf. [Plotkin 70], [Plotkin 71]) on inductive generalization which now serves as a basis for Inductive Logic Programming [Muggleton 92]. See also the more pragmatic approaches such as those of Vere [Vere 80], Hayes-Roth and McDermott [Hayes-Roth & McDermott 78], Kodratoff and Ganascia [Kodratoff & Ganascia 86]...

This paper constitutes an attempt to precise the logical status of the matching mechanism in inductive inference. We shall see that it is closed to Aristotelian induction in applied domains, such as natural sciences or rhetorics. It will then be possible to draw a parallel between Aristotelian biology and heuristics employed in symbolic Machine Learning. To make this parallel meaningful we shall first recall the logical status of induction for Aristotle and other philosophers. We shall then precise what is structural induction, how it can be formalize with inductive logic programming (ILP) and what are the obstacles it faces, before explicating the parallel with applied Aristotelian inductive inference.

Logical Approaches to Induction

The status of induction has remained problematic throughout the history of philosophy and logic. Since Aristotle, many philosophers – e.g. Sir Francis Bacon (1561-1626), John Stuart Mill (1806-1873), Jules Lachelier (1832-1918), Rudolf Carnap (1891-1970), Jean Nicod (1893-1931), Carl Gustav Hempel (1905-1997), Nelson Goodman (1906-1998), Jaakko Hintikka, ... – have tried to show that induction is an inference, i.e. a logical operation they have attempted to legitimate. But these endeavors to legitimate induction through a logical formalism have foundered on a certain number of obstacles. Our purpose is not to enumerate all those

essays and the subsequent obstacles they encounter. It is just to recall the role played by inductive inference in systems of logic and its relative position compared to other inferences, in particular to deduction. For the sake of clarity, we shall restrict here on two main positions, the Aristotelian one, which considers induction as an inversion of deduction, and another, defended for instance by John Stuart Mill, for which the induction is a kind of deductive inference.

Aristotle's Inductive Inference

Although induction probably originated before his time, Aristotle was most certainly one of the first to have spoken of it and to have given it philosophical status, even if its place in his work was not that important. Plato, on the other hand, never mentioned induction. Of course, this was neither an omission on Plato's part nor an idle reference on the part of Aristotle, it was merely a difference in philosophical attitude. For Plato, knowledge is given in the beginning to man who merely has to recall it, to remember it through a process of anamnesis, as is so beautifully illustrated in the Meno dialog, whereas for Aristotle the soul is a copybook on which the world inscribes itself or rather on which we write down the world as it comes to us in the course of our lives.

Syllogistic lies at the heart of Aristotle's approach: the Philosopher analyzed specific modes of reasoning such as refutation, abduction, reductio ad absurdum, petitio principii, etc. in term of syllogisms. It was in this context, which was the subject of Book Two of the *Prior Analytics*, that induction was considered. In order to understand exactly what Aristotle meant by it and the status he gave it, let us examine the definition and example that he gave.

“Induction, then - that is, a deduction from induction - is deducing one extreme to belong to the middle through the other extreme, for example, if B is the middle for A and C, proving A to belong to B by means of C (for this is how we produce inductions). For instance, let A be long-lived, B stand for not having bile, and C stand for a particular long-lived thing, as a man, a horse, or a mule. Now, A belongs to the whole C (for every bileless thing is long-lived); but B (not having bile) belongs to every C. If, then, C converts with B and the middle term does not reach beyond the extreme, then it is necessary for A to belong to B”. [Aristotle, p. 99].

In a word, A corresponds to the term “long-lived”, B the middle of the syllogism to "not having bile" and C to "man, horse and mule", which gives the following figure of reasoning, along the lines of Aristotle.

1. “A belongs to the totality of C”, in other words *All C's are A's*, since men, horses and mules, in other words all animals that are long-lived, live long. This is an observation taken from experience and is posited as a preliminary premise of the induction.

2. “But B also belongs to all C's”, i.e. *All C's are B's*. In other words, another empirical observations tells us that men, horses and mules are all animals not having bile.

3. If we suppose that the middle term B does not reach beyond the major C, i.e. All animals not having bile are men, horses and mules (i.e. since All C's are B's then All B's are C's), then we can infer by induction that All B's are A's, which gives that All animals not having bile live long.
If we use Aristotle's notation, this figure comes from the implicit syllogism below.

All C's are B's	**Major**
All B's are A's	**Minor**
All C's are A's	**Conclusion**

Taking the above example, this gives:

Men, horses and mules do not have bile	**Major**
All animals not having bile live long	**Minor**
Men, horses and mules live long	**Conclusion**

Thus, for Aristotle induction moves from the Conclusion and the Major, which have been shown empirically to hold perfectly, to the Minor. To put it another way, starting with two propositions, *Men, horses and mules live long* and *Men, horses and mules do not have bile*, induction enables us to infer the minor of the syllogism which links the first two propositions, namely *All animals not having bile live long*. In other words, Aristotelian induction inference is an inversion of deduction, which could be summarized as "Induction = Deduction^{-1}"
Let us note that Aristotelian induction appears to be in some ways a certain reasoning. However, in order to be certain, Aristotelian induction requires that the extension of the Major, i.e. C, be covered exhaustively by the extension of the middle, i.e. B, which limits its use to elementary cases. What would happen if the set of all the cases covered by a rule were infinite? Because of this, certain induction is an extreme figure of empirical reasoning that can never be fully realized in practice, since a new case could always appear that would invalidate the existing induction.

Mill's Inductive Syllogism

In the 19th century, many philosophers were interested in induction: in France we find Pierre-Paul Royer-Collard, Victor Cousin and Jules Lachelier, and in England the philosopher who today is considered to be the greatest classical theoretician of induction of them all, John Stuart Mill. In his system of logic he presents induction as being the source of all knowledge.
Once induction has been clearly identified and distinguished from the other modes of reasoning such as *abstraction*, *description* or *colligation*, with which it is often associated, Mill defines it as a formal operation. To do this he uses Aristotle's definition but twists it so that induction becomes a syllogism in its own right and not a mode of reasoning. Remember that for Aristotle induction consisted in looking for one of the premises of a syllogism by starting from the other premise and the

conclusion. Mill transformed this into a syllogism, but of a particular kind. Let Mill explain what he means.

> “[...] every induction is a syllogism with the major premise suppressed; or (as I prefer expressing it) every induction may be thrown into the form of a syllogism, by supplying a major premise. If this be actually done, the principle which we are now considering, that of the uniformity of the course of nature, will appear as the ultimate major premise of all inductions, and will, therefore, stand to all inductions in the relation in which, as has been shown at so much length, the major proposition of a syllogism always stands to the conclusion, not contributing at all to prove it, but being a necessary condition of its being proved [...].” [Mill, Chap. 3, para. 1, p. 345].

In order to understand exactly what is happening, let us take Aristotle's syllogism which was used to illustrate Aristotelian induction.

Men, horses and mules do not have bile **Major**
All animals not having bile live long **Minor**
Men, horses and mules live long **Conclusion**

Remember that for Aristotle induction means moving from the Conclusion and the Major to the Minor.

What Mill does is to transform this syllogism so that the Minor and the Conclusion change places. In this way the proposition that is inductively inferred becomes the conclusion and the initial conclusion replaces the original minor. For the resulting syllogism to remain valid, he adds a new Major that he derives from the principle of uniformity and which says that what is true for men, horses and mules is true for all animals not having bile, which gives:

What is true for men, horses and mules is true for all bileless animals **Major**
Men, horses and mules live long **Minor**
All animals not having bile live long **Conclusion**

This syllogism seems rather strange in comparison with the classical Aristotelian syllogisms seen above, as it is a hypothetico-deductive type of syllogism such as was introduced in Antiquity, after Aristotle, by the Stoics and of which the well-known *Modus Ponens* and *Modus Tolens* are special cases. Having said this, it is well and truly a syllogism, which thus enabled Mill to reduce induction to a conjectural deduction or, more precisely, to a certain deduction under a conjectural hypothesis, namely here the uniformity hypothesis which states that “What is true for man, horse and mule is true for all animals not having bile”.

Symbolic Machine Learning

Symbolic versus Numeric

It is common to regard traditional artificial intelligence as being restricted to symbolic, exact and deterministic approaches, whereas new artificial intelligence would take into account approximation and uncertainty with a numeric approach combining neural networks, belief networks,

reinforcement learning and genetic algorithms, for instance. However, specialists in artificial intelligence do not all identify with this way of looking at things. Knowledge representation is a crucial part of their work and the ontologies to which they sometimes refer are only put forward as hypotheses and models, with nothing definitive or rigid behind them. Therefore, in the last few years nothing that has happened in artificial intelligence and in machine learning seems to reduce the opposition between these different views, even if it is now common to combine symbolic and numeric approaches.

Moreover, this opposition has persisted throughout the history of philosophy and has still not been resolved. Symbols predated numbers and continued after numbers appeared. Numbers were introduced into the theories of induction in the 18th century by Buffon and Reid and have remained, especially in the 20th century, during which the philosophical theories of induction have all had recourse to numbers without forgetting symbols. This was true for the probabilistic theories of people like Carnap and Reichenbach and also for those, including Nicod and Hempel, whose work was an attempt to legitimate a logical approach to induction. This is also true in artificial intelligence where the numerical approaches to machine learning does not signifies the definitive rejection of symbols.

We shall not discuss this opposition between numerical sub-symbolic artificial intelligence and logic oriented symbolic artificial intelligence; our goal is only to see what exactly is new about the inductive mechanisms used in artificial intelligence. However, as we shall see, it appears that symbolic artificial intelligence define a new approach to induction whereas numerical machine learning relies on well-known and well-establish philosophical principles that have already been clearly described by philosophers. More precisely, classical mechanisms on which are based numerical machine learning techniques does not bring anything new. For instance, the induction of association rules by detecting correlation among descriptors, the discrimination of positive and negative examples by finding the optimal separation, the generalization of propositional depiction by dropping one of the conjunct and the introduction of numbers to quantify degree of confirmation of an induced hypotheses largely predate artificial intelligence and machine learning.

This does not mean that the considerable amount of work which has been done in the last few years in numerical approaches to machine learning has not produced anything new. Today's research is more adequate and precise than ever, yet most of the elementary mechanisms which are commonly used in numerical machine learning are based on well-known pre-existing ones. The novelty lies mainly in the way these mechanisms are combined and applied. For instance, the notion of simplicity viewed as controlling the application of generalization operators is nothing other than a case of Occam's Razor principle.

Knowledge representation is a key issue in artificial intelligence in general and, more specifically, in both numerical and symbolical Machine Learning. However, there is a strong difference between the way objects are represented in numerical and symbolical Machine Learning: while representation used in numerical Machine Learning are restricted to sets of features, symbolical Machine Learning is able to deal with structured objects containing subparts that are related each others by logical relationships. It terms of logic, it means that numerical Machine Learning algorithms are trained on examples described by conjunctions of propositions whereas symbolical Machine Learning algorithms authorize first order predicate logic. However, this knowledge representation augmentation, has a mathematical counterpart: in the case of propositional logic, the generalization space algebraic structure is a lattice while in the case of first ordre predicate logic it is far more complex.

To clarify this point, let us take three examples given in propositional logic: e_1 = white $\wedge$ rose; e_2 = yellow $\wedge$ narcissus; e_3 = white $\wedge$ narcissus. The generalization of two conjunctive descriptions retains all common propositional descriptors belonging to both. Therefore, the generalization of e_2 and e_3 is “narcissus”; the generalization of e_1 and e_3 is “white”; the generalization of e_1 and e_2 is empty. More generally, there always exists one maximal common generalization for each pair of examples; the lattice structure of the generalization space is based on this property.

Now, let us consider two flower baskets represented as structured examples:

$flower\_basket_1$ = white(a)$\wedge$rose(a)$\wedge$yellow(b)$\wedge$narcissus(b)$\wedge$on_top(a, b)
$flower\_basket_2$ = white(c)$\wedge$narcissus(c)$\wedge$yellow(d)$\wedge$rose(d)$\wedge$on_top(c, d)

Because those two examples are structured, they refer to predicates, i.e. to functions and not to propositions. As a consequence, descriptions are build on terms which are all different; for instance “white(a)” is not equal to “white(c)” even if they designate a similar property. Therefore, it is not possible to define generalization as an intersection of common conjuncts belonging to descriptions; it is necessary to consider the mappings of their subparts. For instance, one may consider that “a” maps onto “c” and “b” onto “d”, or that “a” maps onto “c” and “b” onto “c”, etc. Since there are two objects in $flower\_basket_1$ and two in $flower\_basket_2$ the total number of map possibilities is $2\times2=4$. Here are the four maximal generalizations corresponding to the four possible matching:

white(X)$\wedge$yellow(Y)$\wedge$on_top(X, Y)
white(X)$\wedge$rose(Y)
yellow(X)$\wedge$narcissus(Y)
rose(X)$\wedge$narcissus(Y)

For more than 35 years now, researchers tried to clearly define the generalization of structured examples with the aim to establish logical foundations of inductive Machine Learning. Different formalisms have been developed by Plotkin, Vere, Michalski, Kodratoff and Ganascia,

Muggleton etc. The next section is dedicated to the presentation of the most wide-spread today, which is the Inductive Logic Programming (ILP) formalism.

Inductive Logic Programming

The ILP formalism is directly related to logic programming and to automatic theorem proving techniques that use the so-called resolution rule. Therefore, to give an account of ILP, one has first to recall what is the resolution rule that serves as basis for deduction procedures. It will then be possible show how, within this framework, induction may be assimilated to an inversion of deduction, as in Aristotelian induction, which naturally leads, since deduction is based on the resolution rule, to an inversion of the resolution rule.

Automatic Deduction with the Resolution Rule

Many automatic proof procedures are based on the resolution rule. In particular it constitutes the basis for most logic programming languages developed in the seventies or in the eighties, for instance PROLOG. The key operation that serves as foundation for the resolution rule is the unification. Here are some definitions of the fundamental notions.

Definition: Two terms t_1 and t_2 are said to be *unifiable* if there exists a substitution σ of the variables of t_1 and t_2 that makes them equal, i.e. such that $t_1\sigma=t_2\sigma$. σ is called a *unifier* of t_1 and t_2.

Theorem: if terms t_1 and t_2 are unifiable then there exists a *most general unifier* (mgu) σ of those two terms, i.e. a unifier σ such that for all unifier η there exists a substitution θ with $\eta = \sigma\theta$.

Once unification has been defined, it is possible to define the resolution rule.

Definition:

let us consider two clauses, C_1 and C_2, i.e. two disjunctions of literals,

let us suppose that L_1 belongs to C_1, i.e. that L_1 is one of the disjuncts of C_1 and that L_2 belongs to C_2.

If L_1 and $\neg L_2$ (or $\neg L_1$ and L_2) are unifiable with the most general unifier σ, then the resolvent C of C_1 and C_2 by L_1 and L_2 – noted res(C_1, C_2; L_1, L_2) – is define by:

$C = res(C_1, C_2; L_1, L_2) = \{C_1\sigma \,\square\, L_1\sigma\} \cup \{C_2\sigma \,\square\, L_2\sigma\}$

Notation: S being a set of clauses, the derivation of the clause C from the application of the resolution rule to clauses of S is noted $S\ `_{res}\ C$

Theorem: S being a set of clauses, S is unsatisfiable if and only if $S`_{res}\cdot$ where $\cdot$ corresponds to the empty clause, i.e. to the falsity.

Corollary: S being a set of clauses and C being a clause, $S\ `\ C$ if and only if $S \cup \neg C\ `_{res}\cdot$

Inversion of Resolution

Following the Aristotelian conception in which induction is an inversion of deduction, inductive inference is formally defined as an inversion of deductive inference. Since the resolution rule plays a key role in deductive inference, researchers tried to inverse resolution. First attempts were done early in the seventies by G. Plotkin who inversed unification. Twenty years latter, in the nineties, S. Muggleton proposed to inverse resolution. The induction is formally defined as follows.
Being given:
A set of observations o that are supposed to be expressed under the form of a set of clauses L_o.
Background knowledge expressed as a theory θ that does not explain the observations, i.e. such that $\forall o \in L_o$:[θ ` o]
The induction consists in finding a hypothesis α which explains all the observations o belonging to L_o, i.e. such that $\forall o \in L_o$ α ^ θ ` o (see [Muggleton 92]).
Because the resolution rule is complete, it means that $\forall o \in L_o$ α ^ θ `$_{res}$ o.
Since L_o and θ are initially given, this is equivalent to α ^ L_o `$_{res}^{-1}$ $\theta$ where `$_{res}^{-1}$ designates the inversion of the resolution rule.
Without going into detail, inverting resolution rule is a nondeterminate operation that requires to inverse substitutions, i.e. to associate the same variable to different constants that are supposed to be matched. Therefore, the number of possible inductions is directly related to the number of possible matches, which may be huge.
The recent advances in relational learning and Inductive Logic Programming attempt to limit the number of mapping by introducing strong formal constraints. The notions of determinism, of ij-determinism [Muggleton & Feng 92], of k-locality [Cohen 94], of 1-determinacy [Cohen 93] and of structured clauses [Zucker & Ganascia 98] are examples of such restrictions whenever they are formalized in the Inductive Logic Programming framework. Other restrictions have been expressed in other frameworks, for instance the number of conjuncts in rules [Michalski 83] or the notion of "morion" [Zucker & Ganascia 96].
Some of those constraints that restrict the number of mappings correspond to syntactical limitations of the learned clauses; other constraints refers to outside knowledge that authorize some mapping and prohibits other.
Consequently, it appears that no existing formal theory of inductive inference can fully account for the structural induction as it is developed in artificial intelligence, since they just consider that induction is an information contraction, while it is also related to a mapping operation. Therefore, relational learning and inductive logic programming put the emphasis on some aspects of inductive inference, that have been largely

ignored before, while they correspond to a general operation currently achieved in most practical inductive inferences.

Back to the Aristotelian Biology

Despite its relative novelty in the theory of induction, the notion of mapping has already been investigated by ancient philosophers. Aristotle, for instance, in the introduction to his Zoology entitled "Parts of Animals" [Aristotle b], established a correspondence between organs that are all involved in the same biological function.
To take just one example of a biological function, that of locomotion, Aristotle maps birds' wings, fishes' fins and mammals' legs since wings, fins and legs are all involved in locomotion. Aristotle argues that this matching or mapping between functional parts helps zoologists to reduce the effort required when trying to understand the organization of unknown animals by reusing part of the work that has already been done. Thus, zoologists who know how biological functions such as locomotion, perception or reproduction are performed for classes of animals will be able to classify new animals by observing their similarity to known ones, and to understand their organization without having to investigate them fully.
However, even though Aristotle recognized the role of mapping as being central to zoology, he never related it to logic. More precisely, Aristotle's theory of induction which was presented in his logic [Aristotle a], is not at all related to matching, but to the inversion of a deductive syllogism. In other words, it appears that the inductive inference, which is practiced in Aristotle's natural science, refers to matching among subparts of objects, while the inductive logic does not.
Curiously, the situation seems to be quite similar today in artificial intelligence, since many machine learning techniques based on an inductive process do not refer to mapping, which is seen as being beyond the scope of the domain. It is the case for neural networks, belief networks, reinforcement learning, genetic algorithms, etc. On the other hand, both research into structural matching operations and recent advances in Inductive Logic Programming show that matching is a crucial issue and that strong constraints are required to limit the number of possible mappings. It also happens that solutions required to decrease the number of mapping is similar to the Aristotelian solution; it is to provide some a priori knowledge about the function of each part or subpart of a scene and to restrict matching to parts that realize the same function.
As a conclusion, artificial intelligence leads to revisit classical theories of induction in a new way which has not be theorized before, even if it has been anticipated for a long time in empirical practice of induction.

References

Aristotle a : *Prior Analytics*, Hackett, USA, 1989, Translated by Robin Smith

Aristotle b : Parts of Animals, Movements of Animals, Progression of Animals, 1968, Havard University Press.

Carnap R., 1962, Logical Foundations of Probability, 2d ed., Chicago.

Cohen W., 1993, Learnability of Restricted Logic Programs, in IJCAI Workshop on ILP, Chambery, France.

Cohen W., 1994, PAC-Learning Nondeterminate Clauses, in Twelfth National Conference on Artificial Intelligence.

Gold E., "Language Identification in the Limit", *Information and Control, 10,* pp. 447-474, 1967.

Hayes-Roth F., McDermott J., 1978, "An interference Matching Technique for inducing Abstractions", CACM, pp. 401-411.

Kodratoff Y., Ganascia J.G., 1986, "Improving the generalization step in Learning" in *Machine Learning: An Artificial Intelligence Approach, Volume II*, R.S. Michalski, J. G. Carbonell, T. M. Mitchell (Eds.), Morgan Kaufmann, pp 215-244.

Michalski R. S., 1987, *Concept Learning*. In Encyclopedia of Artificial Intelligence. S.C. Shapiro (ed.), John Wiley, Chicester : 185-194.

Michalski R.S., 1983 "A Theory and Methodology of Inductive Learning" in *Machine Learning: An Artificial Intelligence Approach*, R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), Morgan Kaufmann.

Muggleton S. & Feng C., 1992, Efficient Induction of Logic Programs, in Inductive Logic Programming, Muggleton S. (ed.), Harcourt Brace Jovanovich, London.

Muggleton S., 1992, *Inductive Logic Programming*. In Inductive Logic Programming, S. Muggleton (eds) Academic Press, London 3-27.

Plotkin G., 1970, "A note on inductive generalization", in *Machine intelligence* 5, pp. 153-163, eds. Meltzer B. & Michie D., Edinburgh University Press.

Plotkin G., 1971, "A further note on inductive generalization", in *Machine intelligence* 6, pp. 101-124, eds. Meltzer B. & Michie D., Edinburgh University Press.

Turing A., 1950, "Computing Machinery and Intelligence". *Mind 49:* 433-460.

Valiant L., 1984, "A theory of the learnable", Communications of the ACM 27, 1134-1142.

Vapnik V., 1995, The Nature of Statistical Learning Theory, Springer.

Vere S. A., 1980, "Multilevel Counterfactuals for generalizations of relational concepts and productions", Artificial Intelligence, Vol. 14, pp. 139-164.

Zucker J-D & Ganascia J-G, 1996, Changes of Representation for Efficient Learning in Structural Domains, in proc. of ICML (International Conference in Machine Learning), Bari, Italy, Morgan Kaufmann.

Zucker J-D & Ganascia J-G, 1998, Learning Structurally Inderminate Clauses, in Proc. ILP-98, Madison, Wisconsin, USA, Page D. (ed.), Springer.

AN INTERACTIVE TREATMENT OF TRUTH : THE DIALOGICAL LOGIC OF VERIDICITY

DENIS VERNANT

Department of philosophy, University of Grenoble II, Denis.Vernant@upmf-grenoble.fr

The question of truth is central to the Philosophy of Language. Only one proposition can work or not work for a given a fact, and for the same fact only one propositional inference may be valid or not. Truth has always been dealt with in a purely monological way, according the tradition in Logic. For Aristotle, a proposition is true if it corresponds to a fact[43]. Stoicism holds that reasoning is valid if and only if it adheres to the canonical form of a trope. But this means forgetting that Truth is always the result of an often long and complex research process relying upon enquiry, questioning or in short, dialogue. The Megarics they think any usage of rationality is "dialectical" which means dialogical and Frege recalls quite relevantly that all knowledge is constructed as a response to a question.[44]

We would like to pinpoint the dialogical aspects of the search for truth and propose a model for the dialogical logic of truthfulness that not only shows the « material » dimension of truth but also the « formal » dimension of validity.[45] The model exploits recent developments in

43. Cf. Aristotle (1969 : 14b15-22, 70). This is the origin of the correspondantist definition of Truth. We may also take note of the fact that the Aristotelan dialetic allows for an explicitely dialogical dimension in argumentation, cf. Aristote (1965 : 313-368).

44. Cf. Frege (1971 : 176).

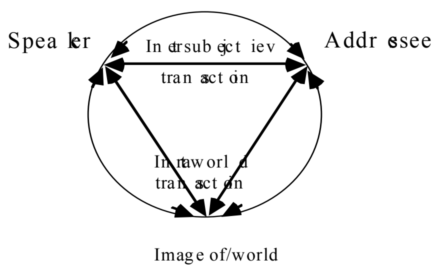
45. The logical distinction between formal/material go back to Aristotle. Here “material” means “contentual”. So, arithmetical truth is in this sense “material” in so far as arithmetics is a mathematical theory which deal with a world – very abstract indeed – the one of the integers, cf. infra, § 3.2.1.

dialogical logic[46] and is implementable on computers as well as allowing for normative usage that provides proper rules so that rational agents may search for truths together and descriptive usage that allows one to analyse real dialogues between human agents.

4. 1. The dialogism of veridicity

1.1 The dual dimension of dialogue

I will deal with both the interactional and the transactional dimension of dialogue [47] :



First of all, a dialogue is a *discursive interaction* which unfolds in an unforeseeable process resulting from a co-operation between at least two interlocutors who interact while simultaneously implementing projective dialogue models[48]. This interaction does not have its finality in itself. It is *heteronomous* and stems from transactional finalities which are intersubjective and intra-worldly. Generally, a speaker does not talk in order to talk, but talks for, with, or against an interlocutor with the purpose of acting on the world that they jointly construct. The *intersubjective transaction* is the movement by which interlocutors recognize themselves mutually as co-speakers in their psychological, social, ideological, etc. dimensions. *Intra-world transactions* question the relation of co-agents on the problem that they encounter in a shared situation.

1.2 The twofold dimension of veridicity : validity & truth

---

46. For a reminder of standard dialogical logics, cf. Vernant (2001, § 1.3.3; 2.4.3; 3.3.3). In what follows I exploit the potentialities of “indoor” games (Lorenzen and Lorenz, 1978) as well as those of “outdoor”games (Hintikka, 1985), cf. Vernant (2001, 2.4.3; 3.3.3).

47. In this diagram, the bows symbolize the interactions, the segments the transactions. For a definition of dialogue in terms of situated and joint activity, cf. Vernant (1997 : 87-107).

48. I put forward a projective model of informational dialogue in Vernant (1997 : 107-126).

Within this theoretical framework, a first distinction has to be made between veracity and veridicity. *Veracity* is a matter of saying what one believes to be true; here, the relevant contrast is that found between sincerity and lying. It concerns the faithful or deceptive *expression* of the speaker's belief, whether that belief is true or false. Only the intersubjective dimension is concerned. This dimension is that of seduction and/or manipulation. In the extreme case, it is simply a matter of convincing an interlocutor. In such a case, arguing is pure rhetoric and the rules of the dialogue pertain to the art of persuasion, *dialectics*[49]. This dialectic dimension proves to be crucial in social life as in the political arena, and it is impossible to separate in practice veracity and veridicity.

In this paper I will concentrate solely on the question of veridicity and therefore proceed on the assumption that interlocutors are sincere in there enonciations [50]. However, I shall examine later another form of veracity which takes place at the level of the dialogical game itself and of the strategic choices of the players (cf. *infra*, § 4.2).

Veridicity puts truth directly into play in its dialogical dimension, in that it results from an agreement between interlocutors at the end of a process of interaction: the speakers agree dialogically to hold a proposition to be veridical.

This first of all presupposes that the interlocutors recognize in each other a minimal degree of logical expertise, which allows them to argue reasonably and possibly to discover the logical inconsistencies of their reasonings, the contradictions between their assertions. This means that the interlocutors share logical capacities of deductive reasoning. We are then in the domain of formal *validity*.

However, the veridictional nature of dialogue must not lead to its confusion with pure logical games. In its praxeological finality, dialogue must bring solution to a *problem* (*Aufgabe*), an actual difficulty in a given situation. Thus, the question becomes that of the material *truth* of the propositions describing the situation at stake. To the requirement of proof for propositional sequences must be added something of the nature of a test: a *procedure of investigation* which allows interlocutors to agree on the truth value of atomic propositions set out[51]. The nature of this

49. Modelization of this kind can be found in D.N. Walton & E.C.W Krabbe (1995); F.Van Eemeren & R. Grootendortst (1996). The logic/dialectic distinction goes back to Aristotle; today, it is expressed in terms of the relations between formal/informal.

50 In Vernant (1997: 59-86), I treated lies as the obverse of assertion and have crossed the dimensions of veracity and veridicity.

51 Hintikka's semantic games (1973) opportunely remind us of the need for verification. (His mistake is to reduce the external verification procedure to a simple dialogue with Nature conceived as an objector.) What Hintikka calls Nature in fact must function dialogically like a third party; cf. Vernant (2004).

procedure changes with the objects in question and the goal considered. One does not solve a problem in mathematics as one does a problem in physics, a moral dilemma in the same way as an daily difficulty, etc. But in all cases, the agreement of interlocutors presupposes recourse to a *third* party jointly admitted as an authority, an indisputable reference. This can be a mathematical theorem, the result of an experimental protocol in physics, the reading of a sacred text or a recipe book, the consulting of a dictionary or data base, call for an expert, etc.

Veridicity is the result of a *dialogical agreement* which presupposes, at the interactional level, that the interlocutors recognize their mutual consistency and at the transactional level, that they mutually share the judgment of a third party which testifies to the truth of atomic propositions bearing on the world at issue. It is therefore important to construct a system of dialogue that allows appreciation of the logical, formal validity of reasoning as well as the material truth of the atomic propositions in question by means of an appeal to external verification procedures recognized as independent judges.

*5. 2. A Dialogical Logic of Veridicity (***DLV***)*

I propose to give an account of the argumentative dimension of dialogue by using the standard logic in its functions of proof theory and model theory.

The Dialogical Logic of Veridicity presents a game's form defined by local rules for logical operators, by global dialogical rules specifying the roles and functions of the players ; by strategic rules of play aand valuative procedures for propositions.

5.1.1 *2.1* Rules of use for logical operators

Rules of use govern dialogical functioning of the operators of standard logic.

RU1 – Negation: If one of the players puts forward $\neg A$, the other attacks with A. Then, there is no possible defence.

RU2 – Conjunction: If one of the players puts forward $A \circ B$, the other can attack by questioning the first conjunct (?1), then the second (?2) of the conjoined propositions. Then, the first player must defend these two propositions.

RU3 – Disjunction: If one of the players puts forward A v B, the other can attack by questioning this disjonction. The first player must defend one of the two disjuncts.

RU4 – Conditional: If one of the players puts forward $A \supset B$, the other attacks by asserting the antecedent A. If the first player cannot reject this assertion, then he must defend the consequent B.

RU5 – Universal quantification: if one of the players puts forward $(x)Fx$, the other can attack by asking what the case is concerning any value a ($?a$). Then the first player must establish the proposition Fa.

RU6 – Existential quantification: if one of the players puts forward $\exists xFx$, the other can attack by asking the first player to provide him with an example (?). Then the first player must establish the proposition Fb.

5.1.2 *2.2* Dialogic rules

Dialogical rules condition the general functioning of this finite zero-sum game with complete and perfect information[52] while stipulating the dialogical functions of the players: Proponent/Opponent; the authorized moves: initial proposition, attack/defence; the procedures of valuation: commitment on/consideration of an atomic proposition.

DR0 – The dialogical game develops by moves which are put forth alternately by an opponent (***O***) and a proponent (***P***).

DR1 – The proponent opens the game by asserting an initial compound proposition. Each player can then choose to attack (or counter-attack) (***A***) or to defend (***D***).

DR2 – If several attacks are produced, the proponent can respond by the attack of his/her choice (including an attack to which s/he has already responded while *reviewing* his/her defence).

– For the formal dimension of the game (validity):

DR3 – The proponent cannot *introduce* an *atomic* proposition that *has not already been asserted* (***AS***) *by the opponent* (and s/he cannot review his/her defence with the help of an atomic proposition unless, again, it has been introduced by the opponent).

– For the material dimension of the game (truth):

52. The game can become a complete and imperfect information game when mixed quantifiers intervene, cf. the Independence-Friendly Logic of Hintikka (1973) and Vernant (2001, 331-3).

DR4 – The interlocutors (Proponent/Opponent) cannot have a *shared commitment* (***SM***) about an atomic proposition unless there is mutual agreement already as to its *valuation*.

DR5 – The interlocutors cannot have a *shared consideration* ***(SN)*** about an atomic proposition unless there is already mutual agreement to admit it as a mere *hypothesis*.

Remark: In the formal dimension of the game, in which alone the opponent can introduce atomic propositions, the purely formal agonistic game does not admit a third party [which is what the central line separating the protagonists symbolizes][53]. On the other hand, the material dimension of the game presupposes calling on a third party which allows the interlocutors to agree on the truth value of atomic propositions [symbolized by a central column which represents the mutual agreement] (see *infra*, § 2.4).

2.3 *Strategic rules*

SR1 – of relevance:

In cases where a strategic choice becomes available, the *relevant choice* is the one that *maximizes* the possibilities of gain.

SR2 – of correction:

Correction governs *respect* of the rules of the dialogical game as well as the associated procedures of investigation providing the valuation of atomic propositions.

SR3 – of decision:

The loser is the player that does not *want* to or *cannot* advance any further arguments[54]. Then, the adversary wins.

Commentary: Strategic choices open up with an attack on a conjunction; with the defence of a disjunction; with defence against an attack on a conditional (response to the attack or defence); with attack on a universal; and with the defence of an existential quantifier. In the formal dimension of the game, relevance consists in bringing the opponent to concede a maximum of atomic propositions that one will then be able to use against him/her. In the material dimension of the game, relevance amounts to making the opponent admit a maximum of the true atomic propositions that are introduced by the proponent.

53. Dialogical logic thus defines validity regardless of the truth of the propositions in question, cf. Lorenz (2001, 258).

54. This allows cases where the interlocutor decides to lose the game deliberately by making the wrong strategic choices, cf. infra, dialogue 5.

2.4 *Rules of valuation*

– Introduction of atomic propositions in a “material” game:

RE1 – By shared commitment resulting from an investigation procedure relative to the world in stake, an atomic proposition can be introduced in the dialogue as *admitted* (***SA***), *rejected* (***SR***) or even of unknown value (***SU***)[55].

RE2 – By mutual agreement an atomic proposition can be introduced in the dialogue as merely a *considered* (***SN***) *hypothesis* [indicated in the central column by square brackets]. It then opens a hypothetical *subdialogue*[56].

– Veridicity of complex propositions:

RE3 – ∃-veridicity:

A complex proposition is ∃-*veridical* if the proponent that assserts it initially wins the play in question.

RE4 – U-veridicity:

A complex proposition is *U-veridical* if the proponent that assserts it initially while respecting the rules of the game (those of relevance and correction), win all games, *i.e.* if s/he wins regardless of the opponent's choices. The player then possesses a *formal winning strategy*.

Commentary:

In the “material game”, atomic propositions introduced by each interlocutor must have been verified by a transactional procedure of investigation jointly accepted by the two interlocutors [concretely, this third party is represented by a central column stating the valuated statut of all introduced atomic propositions : ***SA***, ***SR***, ***SU*** or ***SN***].

∃-veridicity of complex propositions impugns the result of a *particular* dialogue. This result, which is dependent of the idiosyncratic choices made by the interlocutors during a given play, is purely

55. The system is regulated by the principle of bivalence. When the interlocutors admit that they cannot know the truth value of an atomic proposition, it remains open. In this case, the proposition at stake can be simply considered as a mere hypothesis.

56. An exemple is proposed in Vernant, (2004 ; 108). This type of subdialogue is used in each reductio ad absurdum. Ta admit an atomic proposition in a material game is for the two players jointly be committed on its truth relatively to a particular investigation procedure (and vice versa for the rejection). To these joint dialogical acts (admitting, rejection and considering) correspond the illocutionary forces of assertion, denegation and consideration, cf. Vernant, (2003).

contingent. In this case, the players are persons considered in all their complexity of embodied and situated agents.

On the other hand, U-veridicity establishes a valid result *universally* insofar as it can be confirmed by any pair of interlocutors who observe the rules throughout and make always the relevant choices. Logical validity depends on this U-veridicity insofar as a compound proposition is established being true in all possible worlds. In this case, the players are purely rational and act as relevance maximizers which respects all the game's rules and the procedures of investigation.

2.5 *Representation of a play*

For informatic implementation of our model, every move of a play must be precisely described. In addition to the compound proposition introduced, each move is caracterised by a series composed of the number corresponding of the talking turn, the letter specifying the function of the speaker (**O**pponent/**P**roponent), the strategic-type (**A**ttack/**D**efence) and the proposition aimed or the move stated[57]. In the special case of introduction of an atomic proposition in an "material" play, the valuating status take the place of the strategic-type: (***SA***, ***SR***, ***SU*** or ***SN***) and the mutual agreement is expressed be the fact that the speaker's function is assumed jointly by : Opponent + Proponent (***O***+***P***).

For instance, the first move can be noted : "(*p* v *p*), < 1, ***P***, ***AS***, 0>". The joint admission of the atomic proposition *q* is noted for exemple by : "*q*, <*n*, ***O***+***P***, ***SA***, *m*>". So "¬*r*, <*m*, ***O***, ***A***, *n*>" expresses that at *m* the opponent attacks the proposition introduced by the proponent at *n*.

A *round* is a dialogical sequence which oppens and ends by two symmetric moves, typically an attack and the corresponding defence, for instance : *A*, <*m*, ***O***, ***A***, *n*> and *B*, <*m* + *x*, ***P***, ***D***, *m*> [note that *m* is odd and *n* even].

3 The two dimensions of veridicity games

3.1 The formal games

In order to distinguish correctly these formal and material dimensions of veridicity, consider the simple case of a purely formal, ideal game, one in which the stake is logical validity of a compound

57. For instance, in dialogue 1 below, in 6 the move is the attack of the disjonction (?) by the opponent.

proposition. Suppose a proponent who must defend the initial proposition : $\neg(p \vee q) \supset \neg p$

Consider the following dialogue [attacks are in bold type]:

	O	*P*
1		$\neg(p \vee q) \supset \neg p$
2	$\boldsymbol{\neg(p \vee q)}$	
3		$\neg p$
4	$\boldsymbol{p}$	
5		$\boldsymbol{p \vee q}$
6	**?**	
7		p [4]

Dialogue 1

This dialogue can be described so :

$\neg(p \vee q) \supset \neg p$, <1, ***P***, ***AS***, 0> The proponent asserts an initial compound proposition.

$\neg(p \vee q)$, <2, ***O***, ***A***, 1> The opponent attacks the conditional by its antecedent.

$\neg p$, <3, ***P***, ***A***,1> The proponent defends the conditional by its consequent.

p, <4, ***O***, ***A***, 3> The opponent attacks denied proposition by asserting the corresponding atomic proposition.

$p \vee q$ <5, ***P***, ***A***, 2> The Proponent attacks the denied disjonction in 2 by asserting the corresponding affirmative proposition.

?, <6, ***O***, ***A***, 5> The opponent attacks by questioning the disjonction.

p, <7, ***P***, ***D***, 5> The proponent defends the disjonction by asserting the disjonct p which was *previously* asserted by the opponent in 4.

So s/he wins the game because the opponent cannot counter-attack. Thus, by applying the rules correctly and making the relevant choices, the proponent has a *winning strategy* which establishes the *validity*, that is the formal and *a priori* truth of his/her initial compound proposition.

3.2 The "material" games

The aim of a “material” game is to construct dialogically propositions about facts of a shared world. In this dimension, as stipulated by the rules of valuation, we must distinguish two cases.

3.2.1 The ∃-veridicity

In the first case, the veridicity is established by a particular procedure of investigation for one particular world. Take a schematic example by using a purely arithmetic world.

Consider, for example, the arithmetic micro-world made up of integers 1, 2, 3, 4, and the relations $x < y$ and $x = y$. This world, reductible to the set of facts which satisfy the relation at stake: 1 = 1, 2 = 2, 3 = 3, 4 = 4 et 1 < 2 < 3 < 4, can decide between the players of the game. Suppose that an proponent asserts the initial (x)¡y(x £ y). s/he could defend the corresponding disjonction so :

	Opponent	*N*	*Proponent*
1			(x)¡y[(x < y) v (x = y)]
2	**?1**		
3			¡y[(1 < y) v (1 = y)]
4	**?**		
5			(1 < 2) v (1 = 2)
6	**?**		
7		***SA***	(1 < 2)

Dialogue 2

In 1 the proponent asserts the initial disjonction. In 2 the opponent attacks by a universal instantiation on x. The proponent must assert the disjonction for the value 1. In 3 the opponent attacks by asking an instance for y. The proponent chooses 2. In 4 the opponent attacks the disjonction so instantiated. The proponent chooses one of the two disjoints: 1 < 2. In the micro-world considered *this* proposition is true and must be mutually admitted. Then the proponent wins the play.

In this pedagogical example, the world considered is the abstract micro-world of the arithmetic, but naturally the “material” games concern first the “actual” world : the world of physics or the prosaïc world of real life with all its praxeological constraints[58].

3.2.2 The U-veridicity

The logical truth, or formal validity can nevertheless be obtain as truth in all the possible worlds. So we find again the “formal” games by testing proposition not by reference to the actual world, but by

58. For an exemple, see infra, Dialogue 5.

considering *a priori* all the possibles worlds. This is the U-validity, corresponding for instance to general empirical laws[59].

Consider the proposition : $(x)Fx$ 2¡xFx

	Opponent	*Proponent*
1		$(x)Fx$ 2¡xFx
2	$(x)Fx$	
3		?a
4	Fa	
5		¡xFx [1]
6	?	
7		Fa [4]

Dialogue 3

The proponent asserts a conditional. In 2 the opponent attacks by the antecedent. In 3 the proponent attacks this antecedent by asking an instantiation for *a*. In 4 the opponent instantiates with *a*. The proponent defends the initial conditional by asserting the consequent. In 6 the opponent asks for an existential instantiation. In 7 the proponent instantiates with the value *a* by the atomic proposition previously asserted by the opponent. So the proponent wins. The initial proposition is thus U-veridical, *i.e.* truth for all possible world.

4 The descriptive use of *DLV*

All the precedent games, "formal" or "material", come within a *normative use* of *DLV*. Purely rational, the players respect all the rules and make the relevant strategic choises. But this is rarely the case of effective dialogues when the protagonists are faced with complexes situations, serious problems and high stakes. Our model of *DLV* can also deal with these forms of games and may have also a *descriptive use*.

In these effective dialogues, the issues of error and lie recurs. We have evocated the question of veracity for enonciations (cf. *supra*, § 1.2), but this question recurs at the level of the dialogical game itself. A player can want to cooperate and play the veridicity game. In this situation, s/he can either playing *correctly* in respecting all the rules, either *incorrectly* because s/he is tired, distracted, incompetent, etc. In these two cases, s/he is a *veracious* player. But a player can also be *not* veracious. As the rule authorizes (cf. *supra*, § 2.3, *SR3*), s/he can decide intentionally to lose while s/he disposes of a strategy to win the play or even a winning strategy. In this case, s/he *cheats* for instance for make the opponent win (or in all other intention). Or, as in the case of lying in asserting a

59. Cf. infra, § 4.2 the biological law : "It's boys as well girls" in the dialogue between the geneticist and the patient.

proposition, the player can also fail in his/her cheating strategy. So, in this level, four cases are also possible, such that :

VERACITY GAME

Rules / *player*	Correctly	Incorrectly
Veracious	Well played game	Poorly played game
Not veracious	Failure of cheating	Cheat

So, in effective dialogues, we must distinguish the cases when the player inintentionaly plays poorly and the others where s/he intentionally cheats.

4.1 *Poorly played game*

Consider again (cf. *infra,* § 3.1, *Dialogue 1*) a game when the proponent asserts the initial proposition : $\neg(p \vee q) \supset \neg p$. But now, s/he adoptes this strategy :

	Opponent	*Proponent*
1		$\neg(p \vee q) \supset \neg p$
2	$\boldsymbol{\neg(p \vee q)}$	
3		$(p \vee q)$
4	**?**	

Dialogue 4

In 2 the opponent attacks the conditional proposition by putting forth its antecedent. In 3 the proponent chooses to respond to this attack by a counterattack putting forward the disjonction ($p \vee q$). The opponent then attacks this disjonction *and the proponent has no defence* (he does not have the right to put forward an atomic proposition which has not been introduced by the opponent, cf. *DR3*).The proponent has therefore lost *this* particular play while, as we have seen, the proposition is *valide* in so far as there exists a winning strategy to prove it.

In this particular cases, the proponent inintentionaly chooses a *irrelevant* strategy and so doesn't respect the strategical rule *SR1*.

4.2 *Cheat*

But there is also the case when the player doesn't respect the strategic rules even though s/he could do it. Then s/he cheat to win *or even to lose*. Consider this last case which seems the most interesting. It will illustrate one *descriptive use* of our *DLV*.

In this use, our model gives a formal account of real dialogues heard as *effective* processes taken on by two interlocutors and leading to joint decision. Then the dialogical game of veridicity can be played only if the interaction has meaning and finality within an effective transaction between two co-agents and concerns the resolution of a problem in a determinate situation owning its procedures of investigation and action.

Consider the real case analysed by Martine Batt (2003). The corpus is composed of a set of interviews between a patient, Mrs. *P*, on one hand and a geneticist, a neurologist and a psychologist, in predictive medical consultations on the other hand. The patient is requesting a genetic test for Huntington's disease, a neuro-degenerative genetic disease. For my purpose, I shall isolate some dialogical sequences from the interviews between Mrs. *P* and the geneticist, then the neurologist.

In the praxiological sense, the *problem* is that of knowing whether the patient has Huntington's disease by testing her. The stakes of the interview are fundamental for Mrs P. since they consist in her confronting and accepting the verdict.

The first interview, that with the geneticist, consisted in the geneticist's constructing a common representation of the situation while together commenting on the family tree prepared by the patient and specifying each of the family member's medical state. The following micro-world was thus jointly:

: Girls without the disease

: Boys without the disease

: Girls that may be carriers

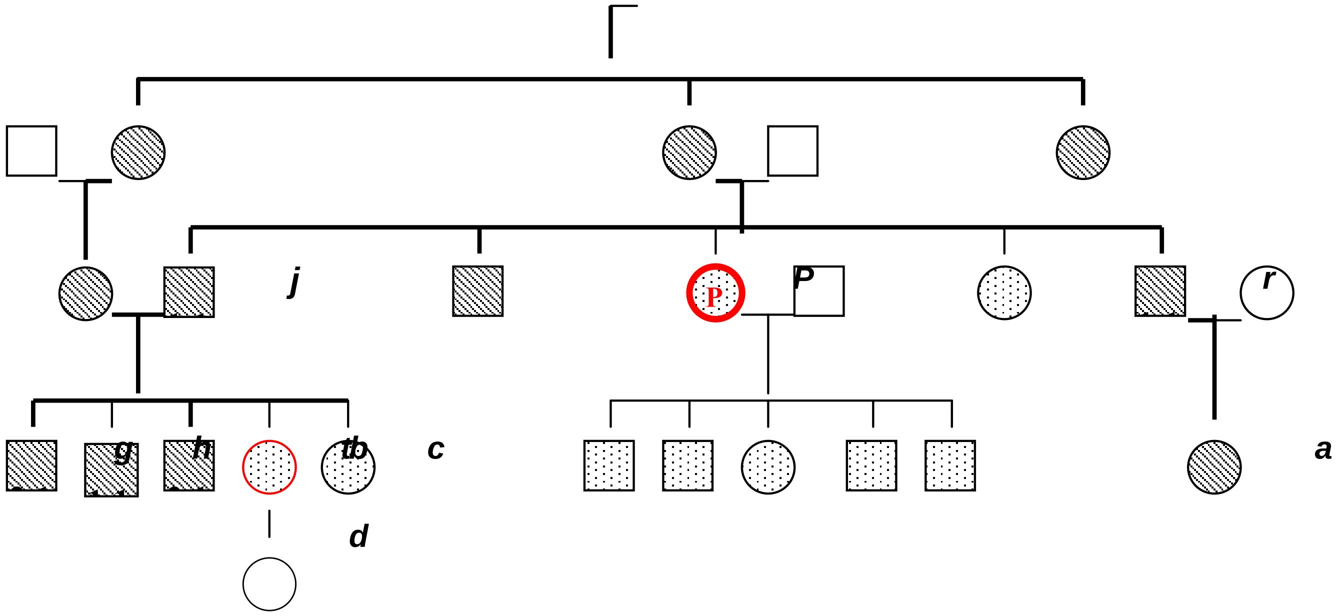
: Boys that may be carriers

: Girls with the disease

: Boys with the disease

: Girls who are not carriers after test

P = Patient ; *j* = Jules ; *r* = Robert ; *a* = Anne ; *b* = Brigitte ; *c* = Corinne ; *d* = Dorothée ; *g* = Ghislain ; *h* = Hervé ; *t* = Thierry



> *5.1.3 The* **investigation** *conducted by the geneticist in the form of a simple informative dialogue allowed the elaboration of a* **common micro-world** *as a reference which, intervening as a jointly admitted third party, could determine the truth value of atomic propositions during the dialogue. For example, by simply reading the family tree, one can learn that Brigitte, the daughter of Jules, is likely to be a carrier of the illness.*

Now consider the following sequence, extracted from the initial dialogue with the geneticist:

P277	a:	they are in Paris,
	b:	then the three boys are ill,
	c:	and they don't work anymore, really in houses
	d:	and the two girls have nothing
G278:		mmh mmh
P278:		they have nothing!
G279	a:	Ah
	b:	but it's not connected to sex, is it,

c: it's boys as well as girls [c'est aussi bien *les* garçons que *les* filles]

d: do you know that?

P279 a: oh, okay

b: no but that's what I'm saying, right

G280 a: yes yes

b: you've seen that the boys have it as well as the girls, right?

[vous avez vu qu'il y a aussi bien *des* garçons que *des* filles qui sont atteints hein]

P280 a: well, yes

b: my brother Robert died, it wasn't a sinecure.

G281 well there you are (*5 seconds*) so Denise, she's Jules' woman, is that it?

The patient introduces the reference to Jules' children in the beginning. She reminds the interlocutor that the three sons have the disease and strongly affirms that the two girls have nothing. Her strategy is no longer aimed only at Brigitte, but also at Corinne. Discerning this strategy, the geneticist corrects her by using a *universal* discourse domain and reminding her of the *biological law* according to which "it's boys as well as girls"[60]. The patient apparently accepts this law which is hard to take ("oh, okay"). The geneticist repeats himself: "you've seen that the boys have it as well as the girls", but he commits the logical error of weakening the universal law [expressed by "les" in French] into existential form [the indefinite article "des"]. The patient can then retract her acceptance ("no but that's what I'm saying") and has no difficulty in showing that some boys are affected.

This dialogue can be represented so [where Bx = x is a Boy ; DHx = x has the disease] :

60. In sound rhetorical form, he should have said: "It is girls as well as boys"; the patient exploits this inversion cleverly by speaking of the boys.

	Geneticist		Patient
1			$(\exists x)(Bx \circ DHx)$
2	**?*n***		
3			$Br \circ DHr$
4	**?1**		
5		***SA***	Br
6	**?2**		
7		***SA***	DHr

Dialogue 5

Adroitly, the patient assumes the role of proponent and confirms the first term of the conjunction introduced by the geneticist (in accordance with the inauspiciously proposed order): $(\exists x)(Bx \circ DHx)$. In 2, the patient puts forward an existential instanciation on Robert, *r*, which gives: $(Br \circ DHr)$. The geneticist must accept this proposition since *Br* was jointly admitted in the micro-world mutually constructed (**SA** *Br* in 5). If he attacks the second term of the conjunction, *DHr*, the patient could assert it equally (**SA** *DHr* in 7). Incontestably, Mrs. *P* has won *this particular* dialogical play. The geneticist ratifies this dialogical victory by a "there you are" and, after a meaningful silence, changes the subject[61] by going back to the construction of the family tree.

This is a clear case of ∃-veridicity: the proponent wins the dialogue but is "wrong" inasmuch as her conclusion cannot be admitted universally. Her ∃-veridicity doesn't imply U-veridicity.

Indeed, the patient, by choosing Robert as her example at the end of the dialogical sequence (and not one of Jules' sons, who were involved at the beginning of the dialogue), changes the world of reference and operates a *referential slip* from the left of the family tree to the right. However, in this new world of reference which is Robert's family, the universal biological law is verified since Robert's girl, Anne, also has the disease. Adopting the same world of reference, the geneticist could comfortably have shown that girls can also be affected. He could have justified the second joint term of the biological law: $(\exists x)(Gx \circ DHx)$ by instantiating precisely on Anne, thus also observing the rule of relevance[62]. The patient would have lost and should have had to face the cruel reality [here Gx = "x is a Girl"]:

61. In fact, the geneticist was not able to or did not want to rid the patient of her deception. During this initial dialogue, he was preoccupied with constructing the medical situation of the patient's family. Besides, he knew that she was subsequently going to have interviews with the neurologist and the psychologist, cf. Batt (2003).

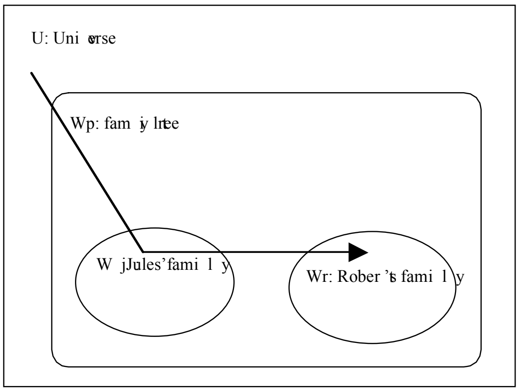
62. Indeed, when a disjunction is proposed, the player must choose to defend the disjunctive component favorable to his /her thesis.

	Patient		*Geneticist*
1			$(\exists x)(Gx \circ DHx)$
2	**?*n***		
3			$Ga \circ DHa$
4	**?1**		
5		***SA***	Ga
6	**?2**		
7		***SA***	DHa

Dialogue 6

U-veridicity would have driven to the geneticist's victory if he had not committed two errors (one rhetorical, the other logical) or if he had imposed the appropriate choice of counter-example, that which concerns Anne. *But He doesn't want to do this and has intentionaly lost the play. He has deliberately made the irrelevant strategic choises.*

One of the interesting things about this dialogue is the subtle game on the reference-worlds that successively plays the role of third party. The biological law expressed by the geneticist has universal value (it is a "law of nature"). In a process of denial[63], the patient first tries to avoid it by deliberately restricting the reference to Jules' family, then by generalizing to Robert's family, of which she deliberately overlooks the fact that the daughter Anne is affected. We see that the status of third party varies in extension during the conversation: from the universe of all individuals subject to biological law to the family tree, then to Jules' family, and finally to Robert's family:



Referential slip

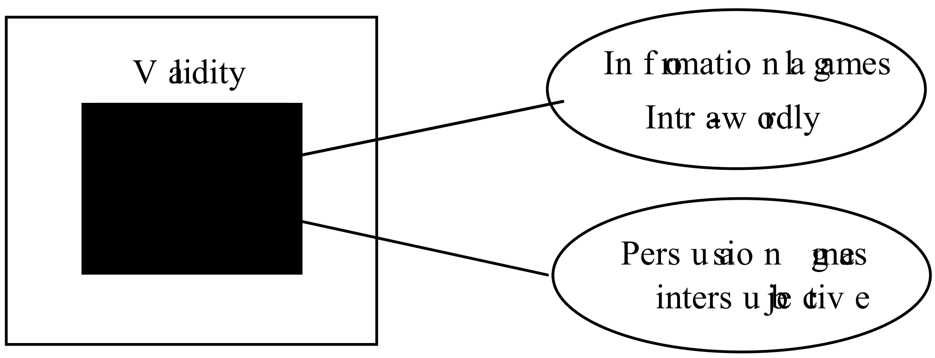
63. On the pragmatic definition of denial, cf. Vernant (2003: 77-90). On this specific case, cf. Trognon & Batt (2004).

6. 4. Conclusion

If the literal meaning of *statements* is provided *in abstracto* by semantics and the sense of *utterances* by pragmatics, the ultimate aim of a discursive interaction resides praxeologically in its transactional, intersubjective and intra-world stakes. As Wittgenstein reminds us: “The meaning of the proposition depends on the rest of our actions”. The same applies to veridictional dialogues. By combining the discursive aspect of validity and the actional aspect of material truth, our model of *Dialogical Logic of Veridicity* aims to account for the two fundamental dimensions of the process of co-evaluating veridicity : “formal” *validity* and “material” *truth*. We have also shown that this model can have two distinct uses : *normative* when the players are agents purely rational logically competent which behave optimally and *descriptive* when the players are actual agents with their rational limitations and their idiosyncrasic impulses, desires, etc.

Our Dialogical Logic of Veridicity that we have put forward clearly remains an ideal game which can only be applied, as we have attempted to show, to *some sequences* of real dialogues. Such dialogues, in their complexity, generally combine several dialogical types of games. For example, in an interaction with cognitive finality, a veridictional sequence can be preceded by an information-seeking sequence and followed by a sequence mobilizing dialectical strategies of persuasion:

RATIONAL GAMES



Furthermore, alongside such dialogues with cognitive finality, dialogues with a different kind of finality have to be admitted – the conative finality pertaining to the dialogues in which negotiations, disputes, etc. are

conducted[64]. My sole objective in this paper has been to give a specific account of the dialogical treatment of veridicity. An abstract logic of propositions held true or false *a priori* should be substituted with a dialogical logic which holds or does not hold a proposition (in the sense of a *proposal*) jointly admitted as a result of a double process of rational argumentation and praxeological investigation.

ACKNOWLEDGMENTS

This research developed out of discussions with Martine Batt and Alain Trognon on the dialogical formalization of sequences from exchanges in predictive medicine. The paper has also benefited from criticisms made by Alain Lecomte and Shahid Rahman. Naturally, I alone am responsible for any remaining errors. I thank Colin Schmidt for his assistance for the drawing up the text and the opportunity of its publication.

Bibliography

ARISTOTE (1965), *Les Topiques*, tr. J. Tricot, Paris, Vrin.

ARISTOTE (1969), *Catégories*, tr. J. Tricot, Paris, Vrin.

BATT M. (2003), *Analyse d'une pratique interlocutoire : la consultation de médecine prédictive, étude d'une consultation prédictive,* thèse sous la direction du Pr. Alain Trognon, Nancy II.

BATT M., TROGNON A. & VERNANT D. (2004), « De la persistance d'une croyance malgré l'interaction », *La croyance en question*, Éric Grillo dir., *Psychologie de l'interaction*, n°17 & 18, Nancy, 117-165.

FREGE G. (1971), « Recherches logiques, la pensée », tr. Cl. Imbert, *Écrits logiques et philosophiques*, Paris, Seuil, 170-195.

HINTIKKA J. (1973), « Quantifiers, Language-Games, Transcendantal Arguments », *Logic and Ontology*, M.K.Munitz ed., New-York U.P., 37-57.

HINTIKKA J. (1985), *The Games of Language*, *Studies in Games-Theoretical Semantics and Its Applications*, Dordrecht, D. Reidel Pub. Comp.

LORENZ K. (2001), « Basic Objectives of Dialogue Logic in Historical Perspective », *Synthese*, 127, 255-263.

LORENZEN P und LORENZ K (1978), *Dialogische Logik*, Darmstadt, Wissenschaftliche Buchgesellschaft.

TROGNON A. & BATT M. (2004), « Dissymétrie des rapports au réel dans un jeu de dialogue de recherche et de découverte en consultation de médecine prédictive de maladie de Huntington », Communication au colloque : *Interactions asymétriques. Communications privées, institutionnelles et médiatiques*, Québec, 26-28 février 2004.

VAN EEMEREN F. & GROOTENDORTST R. (1996), *La Nouvelle dialectique*, tr. fr. Paris, Kimé.

VERNANT D. (1997), *Du Discours à l'action*, Paris, PUF.

VERNANT D. (1999), « Analyses sémantique, pragmatique et praxéologique des phénomènes communicationnels », *Analyse et simulation des conversations. De la théorie des actes de discours aux systèmes multi-agents*, B. Moulin & B. Chaib-draa éds., Lyon, L'Interdisciplinaire, 101-131.

VERNANT D. (2001), *Introduction à la logique standard*, Paris, Flammarion.

64. For a typology of dialogues, cf. Vernant (1999).

parallel session speakers' contributions

Section 1

automaticity communication & thought

INFORMATION SYSTEMS – WHY IT IS A DESIGN SCIENCE

Torkil Clemmensen[1], Jacob Nørbjerg[1], Anirudha Joshi[2]

[1]Department of Informatics, Copenhagen Business School

Howitzvej 60, DK-2000, Frederiksberg. {tc.inf,jan.inf}@cbs.dk

[2]Industrial Design Centre, IIT Bombay

Powai, Mumbai 400076, India. anirudha@iitb.ac.in

Abstract:

In this chapter, we explore patterns in intra- and interdisciplinary relations/conflicts in the software and systems development fields. We examine and compare Software Engineering (SE) and Human-Computer Interaction (HCI) research along two dimensions (1) the nature of the research (normative vs. descriptive); (2) the nature of the research object (the design process vs. the artefact-in-use), and we argue that conflicts within researchers and practitioners from different disciplines are as much related to intra-disciplinary differences in research interests and research object as to gaps between disciplines. We finally suggest that the idea of a design science (Hevner *et al.*, 2004; Simon, 1996) as a way to identify and overcome to resolve intra- and interdisciplinary disagreements in the disciplines that contribute to software and systems development.

Introduction

Software development is fundamentally a multi-faceted discipline. People engaged in producing software and IT systems need to consider not only the program code itself (and how to produce it) but must take into account the purpose of the software, the users and their interaction with software, the complexities of managing the development process itself, technical and economical criteria and constraints, and several other issues. Hence, different disciplines and research areas meet and interact in software and systems development; i.e. organizational science, economy, psychology, programming, software engineering, industrial design, etc. (Hevner et al., 2004; Orlikowski & Barley, 2001). There are, however, significant gaps and disagreements between the disciplines concerning both the nature of the software product, and the process needed to produce it. Is it a computer program, new organizational processes including an information system, the actual interaction between humans and the program, or something else? Recent literature about the gap between for example the disciplines of Human Computer Interaction (HCI) and Software Engineering (SE) illustrates this point (Clemmensen & Nørbjerg, 2003).

The authors of this chapter have backgrounds in psychology, computer science and design, and teach in multidisciplinary degree programs in information systems and design. We experience daily how poor cross-disciplinary understanding leads to disagreements about how each discipline contributes to software and systems development. The computer scientist[65] focuses on the construction of computer programs. He finds interaction design interesting, but perceives it as the *study* of user interaction rather than the *design* of functioning interfaces. The main concern of the psychologist, on the other hand, is how humans interact with computers. He realizes the importance of computer programs but considers them a mere technical necessity, the workings of which (below the interface) are less important to the point of being trivial. The industrial designer strives to be a creative problem-solver who produces, unique and yet practically feasible solutions in an industrial context. Like the psychologist, the designer is concerned with the human users and to him the technology is relevant only to the extent it helps meet design goals.

A closer examination of the different disciplines and the gaps between them reveals, however, strong debates and disagreements *within* the different disciplines about the nature of the design process and the resulting product, as well as interesting similarities between positions or camps *across* disciplines.

In this paper we argue that recognizing the nature of both differences and similarities will help reduce interdisciplinary gaps and disagreements. We further argue that despite the apparent differences and conflicts between disciplines they are all concerned with the same goal: the design of computer artefacts (Simon, 1996); and that the different disciplines and

65 The positions are deliberately exaggerated.

positions within disciplines all contribute to an Information systems design science as discussed in (Hevner et al., 2004; Simon, 1996).

The analysis and discussion focuses on the disciplines of SE and HCI but we believe that the framework and approach used here can be used to examine other disciplines as well.

In the following section we will shortly elaborate on the inter- and intradisciplinary debates in SE and HCI. In section 3 we introduce an analytical framework and use it to identify and describe different research tradition in SE and HCI research. Section 4 discusses the differences and similarities between disciplines and tradition and elaborates on the idea of software and systems development as a design science, and section 5 concludes the chapter.

Inter- and intra-disciplinary debates in SE and HCI

Consider, as a first example, the long standing debate about the nature of programming: As early as 1971 Nicklaus Wirth wrote a paper on program development (Wirth, 1971) in which he proposed a top-down programming process based on systematic decomposition of a problem statement. A number of empirical studies of programming, however, characterizes the process as an evolutionary process where the programmer constantly shifts between reflection of the problem and working with (fragments of) the solution using his experience with programming tools (Guindon, 1990a; P. Naur, 1972; Walz *et al.*, 1993). Thus, the debate concerns whether the design of computer based artefacts is a formal problem solving process, founded in sound engineering principles, or a pragmatic, situated process, defined by the programmer's own experience and emerging and concurrent understanding of the problem at hand and possible solutions.

Secondly, when we discuss the role of psychology in human computer interaction and computer science, it is worth while to distinguish between modern cognitive psychology, with its research focus on the elements of mind, (this tradition can be traced back (Wundt, 1874) and his mental chemistry), and on the other hand much of applied psychology with its focus on the difference an idea or a scientific theory makes in the real world (with traces to the pragmatic psychology of (James, 1890). It is the latter approach that most people (both researchers and practitioners) associate with the psychology of human-computer interaction and usability (Gillan & Bias, 2001), despite early attempts to locate human computer interaction on a basis of a scientific psychology (Clemmensen, 2006). For usability and HCI this means that instead of attempting to control variables such as users' goals and the contexts of actions in order to achieve valid knowledge for a narrow set of circumstances, a pragmatic HCI tradition aims to study those variables as they appear with the aim of making a clear difference in the design of interactive artefacts (Gillan & Bias, 2001).

Thus, there appears to be parallel debates between formalism/structuralism and pragmatism going on in both SE and HCI.

These debates may appear simple, but as researchers and teachers in the disciplines we frequently experience, what we are coming to see as false conflicts between the disciplines arising from confusing the formalism-pragmatism debate with the differences between software engineering, psychology and design. One example of these conflicts is when the computer science teacher presents a curriculum for an introductory course in programming, and the psychologist fails to see the relevance of teaching these low-level technical topics in a university program. The computer scientist however, sees he “simple” programming course as a way to make the students appreciate both the programming process and the nature of the software artefact. Conversely, when the psychologist proposes a user-technology interaction course, the computer scientist misses any trace of development in the curriculum, and thus fails, in the eyes of the psychologist, how an appreciation of the interface goes hand in hand with the ability to build it. In both examples we see how the one side is not paying enough attention to the tacit aspects of the others’ understanding of his own discipline and how it contributes to the overall goal of building computer artefacts.

How can we resolve these differences? Below we suggest that a possible way forward is to build on the idea of an overarching design science (Simon, 1996) that in short says that while what we call “natural” phenomena necessarily evolve due to laws of nature, other phenomena in the world are "artificial" because they depend on the goals of their designer. Putting the goal of the designer – be it software designer, interaction designer or industrial designer – in focus may make the pragmatic / formalist distinctions and the resulting possibilities for category judgment errors across fields less important.

Research traditions in HCI and SE

We have developed the framework in figure 1, to further explore the traditions and their differences within and between disciplines. The choice of the framework’s three dimensions has been guided by our wish to highlight differences and similarities among significant traditions in the disciplines.

1) SE – HCI. This dimension refers to the gap between two important and irreconcilable disciplines in design of computer software. From the point of view of Software Engineering, Human Computer Interaction research focuses on principles of human behaviour and interface design. It does not contribute to systems design and construction. However, from the point of view of Human-Computer Interaction, Software Engineering is preoccupied with the technicalities of systems design and construction but disregards the characteristics and needs of humans.

2) Normative – descriptive. This dimension relates to the philosophical and practical distinction between how reality should be within a given structure of culture and common sense, what is right and wrong, good and bad (normative), and on the other hand, the falsifiable

and positive descriptive explanatory beliefs, theories and explanations of how reality actually is.

3) Artefact in use – design process. This dimension refers in practice to the distinction between product developer companies (who focus on their products, who know what to design) and consultancy companies (who focus on the methods, who know how to design)

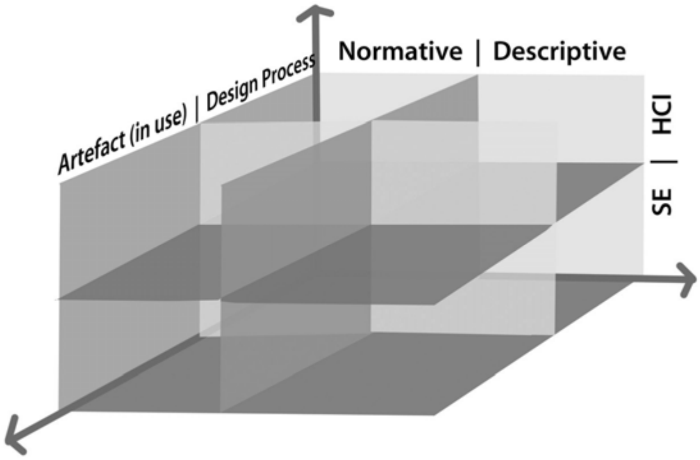
Below, each cell (there are 8) is filled with examples of methods, techniques, approaches that are well known within the disciplines.

Figure 1. Cell framework for understanding different basic positions in information system development research and practice.

Normative software engineering focused on the design process

There is a dominant tradition for focusing on guidelines, methods and process frameworks in Software Engineering research The term software engineering itself was introduced in 1968 (Peter Naur & Randell, 1969) to address the increasing problems with budget overruns and poor quality through a systematic "engineering" approach to the production of software, with clear and well-defined stages, and detailed guidelines for software production (Friedman, 1989).

In the following decades, research on programming techniques,



analysis and design methods, and project management flourished. In what became a classic paper about software project management, (Royce, 1970) recommends to plan and manage the software production process by organizing it as a predefined sequence of stages. Wirth (1971) proposes a "divide-and-conquer" approach to programming, where the programmer systematically breaks the problem into a hierarchy of sub-problems until each sub-problem can be solved by a few simple program statements. Development methods, addressing substantial parts

of the life-cycle such as Structured Analysis and Design (Yourdon, 1989) or Rational Unified Process (Jacobson *et al.*, 1999) integrates a description of the stages in the process (similar to the one provided by Royce) with detailed prescriptions about how to carry out each step, complete with suitable techniques and tools.

Normative software engineering focused on the artefact in use

Normative SE literature focuses much on the process and little on the artefact and its use. (Pressman, 2005) is a seminal book on SE and is valuable, up-to-date reference of current understanding in the field. Though the book summarizes several SE processes, it seems to largely present an inward view of design. If good (building) architecture is all about "firmness, commodity and delight" (Kapor, 1991) SE seems to try to focus on the artefact itself, and to be about firmness (no bugs), but very little about commodity (suitable for purposes) and never about delight (pleasurable experience). Many concerns are related to how we can make sure that we develop software optimally, maintain it well, ensure that it will not crash etc. Few concerns go beyond this to how we can respond to problems of software use.

Descriptive software engineering focused on the design process

A minor research tradition in software engineering has been concerned with the study of actual software design processes. An important driver in this line of research has been a wish to understand how (if at all) the guidelines and recommendations produced by the prescriptive tradition (see section 3.1) are used in practice, resulting in a long stream of research that both adds to and criticizes this tradition and its results (Bansler & Bødker; Madsen *et al.*, 2006; P. Naur, 1972; Stolterman). Others study practice with the aim to understand human design and problem solving activities and capability, either on an individual basis or in a group/organizational setting (Baskerville & Pries-Heje; Fitzgerald; Guindon, 1990a, 1990b; Madsen et al., 2006; P. Naur, 1985; Walz et al., 1993).

This research has added considerably to our understanding of how software is developed in practice, and the important role played by the skills, personal capabilities and experience of the individual software developer, however the research has rarely – if at all – been developed into recommendations for how to develop software or manage software projects.

Descriptive software engineering focused on the artefact in use

In SE research, Agile programming methods while sounding 'process oriented' have been known for their continuous focus on the 'artefact in use', or at least the 'artefact to be delivered'. Agile methods were developed in response to a 'need for an alternative to documentation driven, heavyweight software development processes' (Beck *et al.*, 2001) in an 'effort to overcome perceived and actual weaknesses in

conventional SE processes' (Pressman, 2005). One such weakness is the lack of focus in traditional SE process on the artefact in use (see 3.2 above). Among the 12 principles behind the Agile Manifesto (Beck et al., 2001), 3 are concerned with the artefact in use: 'Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.', 'Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.' and 'Working software is the primary measure of progress.'(Beck et al., 2001).

Normative human-computer interaction focused on the design process

Some of the Human-Computer Interaction literature sees HCI as a kind of system development and tries to set up models for a HCI oriented design process that will tell the HCI professional what is good and bad, right and wrong in user centred design. For example, (Preece *et al.*, 2002) discusses " the process of interaction design", under three headings: 1) activities and key characteristics of the interaction design process, 2) practical issues such as how to generate and choose among alternative designs and 3) lifecycle models that shows how the activities are related. Discussing lifecycle models, (Preece et al., 2002) presents software engineering lifecycle models such as the waterfall model, the spiral lifecycle model and rapid application development side by side with lifecycle models in HCI such as the Star lifecycle model and the usability engineering lifecycle, and conclude that the interaction design process is complementary to lifecycle models from software engineering.

The great focus on normative design processes in human computer interaction is also visible in international standards for practitioners. The ISO 13407 (1999) Human-centred design processes for interactive systems identify four phases in the design process. The four phases constitutes a life cycle of information systems design include 'identify the context of use', 'describe the user requirements', 'create design prototypes' and 'do user based evaluation'[66]. The ideal of HCI research and practice here is both to increase the efficiency of the development process and produce products that improve the work environment for the users of the design artefact.

Normative human-computer interaction focused on the artefact in use

Historically, predictive HCI approaches have relied on norms for artefact use. A famous example is the GOMS family of performance models that predict time to complete a task based on norms for human behaviour and cognitive processes (Card *et al.*, 1983; John, 2003)

However, the main reference on the usability of an information system design in use is the international standard ISO 9241-11 (9241-11, 1998; Jokela *et al.*, 2003) which defines usability as:

> *[The e]xtent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.*

66 See e.g. (http://www.usabilitypartners.se/usability/standards.shtml)

A more detailed account of what is meant by right and good usability is also given in the standard, to explain each of the concepts in the definition. Effectiveness has to do with the user's goals or intended outcome of using the information system design, efficiency with the resources used to reach the goal, satisfaction with the user's attitude towards the product, context is more or less the whole environment and work system is the particular part of the environment used to achieve the goal.

Furthermore, the standard advocates three approaches to identify good usability: 1) analyze the product (is the product user friendly?), 2) analyze the interaction (does the user's mental interaction with the product show signs of usability?), 3) measure directly the components of usability: effectiveness, efficiency and satisfaction (9241-11, 1998).

Descriptive human-computer interaction focused on the design process

Reports on how HCI design is carried out in the 'real' world shows it to be messy and not follow the normative prescriptions from textbooks. For example, (Kelly, 2001b) and (Kelly, 2001a) relates many stories where Ideo broke away from the conventional processes with innovative techniques that had a breakthrough in the real world. As a position to be taken by a member of the design team, this position may be grounded in the designer's extensive experience and anecdotic guidelines for the design process. For example (Mayhew, 1999) suggests a flexible approach in the design process that may be suitable for external consultants to suit the needs of varied products and project teams. She distinguishes between tasks and techniques and recommends that practitioners be flexible with techniques though not compromise on tasks. For example, during design a new conceptual model needs to be evaluated. While the most preferred technique may be evaluation by a user test with users, cheaper alternatives could be remote usability tests and review based techniques such as heuristic evaluation.

Literature also describes what HCI practitioner could do to improve the HCI maturity of an organization. (Mayhew, 1999) suggests a number of techniques that may be useful to get a buy-in for usability from stakeholders and when to 'walk out' of projects. (Mayhew & Bias, 1994) compile a large reference on how practitioners can justify the return on investment in HCI activities. (Battle, 2005) identifies patterns of integration between HCI and SE and lists best practices that can be adopted by other designers such as:'foot in the door for internal usability group', 'foot in the door for external consultants', 'UCD focus on early definition and design' and 'UCD in every phase'. She gives useful advice for HCI practitioners in each of these patterns to help them integrate with SE development process.

Descriptive human-computer interaction focused on the artefact in use

HCI methods are according to some authors always focused on evaluation of the artefact in use (Hartson *et al.*, 2003). Usability evaluation methods are used for formative, qualitative evaluation with the goal of finding lists of usability problems that can be fixed in the iterative design process. It is all about finding qualitative data about the use of the artefact.

Some HCI authors believe, however, that the meaning of usability of artefacts in use – quality in use – is determined by how we measure it quantitatively (Hornbaek, 2006). Quantitative measurement of usability include measures of binary task completion, accuracy measures, recall, completeness, quality of outcome, input rate, mental effort, usage pattern, communication effort, learning measure, preference, ease-of-use, attitudes, perception of outcomes and interaction, and more. This leaves out the analysis of experiences with user interfaces; (Hornbaek, 2006, p81) simply excluded informal usability tests from his review of the current practice in measuring usability.

Discussion

The eight cells in the framework are summarized in table 1 below.

Research area	*Research interest*	*Research object*	*Focus and results*
Software Engineering	Normative	Design process	Life cycle models, guidelines, tools and techniques for *software development*.
		Artefact in Use	Very little focus on the artefact in use
	Descriptive	Design process	Empirical studies of software development Understand how software developers perceive and solve problems.
		Artefact in Use	Focus on deliverable code (agile)
Human Computer Interaction	Normative	Design process	Life cycle models, guidelines, tools and techniques for the *development of human-computer interfaces*.
		Artefact in Use	User interface standards and evaluation criteria.
	Descriptive	Design process	A flexible approach in the design process, suitable for external consultants and varied products and project teams; the return on investment in HCI activities
		Artefact in Use	Studies of the artefact in use. Evaluation of usability based on real world experiments.

Table 1. Summary of the traditions in SE and HCI

The summary shows SE and HCI as different research areas regarding focus, theoretical base, and results. Software engineering is, by and large, concerned with the production of computer software, whereas human computer interaction focuses on the interplay between the software and the human user. Our analysis of the two fields shows, however, that neither of the fields are homogenous, and that the similarities across fields dominate intra-field differences in some cases.

From a software engineering point of view, the most significant difference within, as well as across the disciplines, is between normative and descriptive research. Both SE and HCI have strong research traditions that aim to build life cycle models, methods, guidelines, tools and techniques for the production of software, respectively human computer interfaces. In both fields we also find research that aims to produce product standards and evaluation criteria.

In both fields we also find significant descriptive research which both increases our understanding of important issues in software

development and human computer interaction, but which also assumes a critical stance towards the prescriptive research traditions. Thus, the descriptive research into software development is often motivated by a critical stance vis-à-vis the prescriptive tradition, which, it is argued, has a too simplistic understanding of human design and problem solving activities in general and of the software development process in particular. The descriptive research does not, however, transform the insights obtained into usable guidelines and tools for software development, which renders the results less interesting from the point of view of the prescriptive tradition. From descriptive studies of HCI in organizations we know that user involvement during the design process is in conflict with the actual usability of the designed artefact as experienced by the user; in fact information system design organizations seem to prefer those parts of the user involvement process that they see fitting their own organizational culture (Iivari, 2004).

From an HCI point of view, the most significant difference within, as well as between the disciplines, is between having a focus on the design process or on the artefact in use. Much of what is understood as HCI actually focuses on the artefact in use and some theorists (Hartson, 1998) even argue that having a focus on design methodology means moving outside the boundaries of scientific HCI. In a sense they are very much in line with the traditional view of engineering design: design is a special activity, which concerns itself with ill defined problems and which relies not on the clarity of the principles of how to produce the design artefact, but on the demonstrated appropriateness of that artefact (Roberts *et al.*, 1992). However, as our analysis in this paper has demonstrated, HCI harbours design process theorists, which may have much in common with software engineers.

Does this mean then, that the differences and poor communication between SE and HCI research and practice can be resolved? We would say yes, at least insofar as one recognizes the differences between the normative and descriptive traditions *and* accepts that both contribute towards the design of computer artefacts. Following (Hevner et al., 2004) we will argue that a scientific discipline that is concerned with the design of artefacts needs research that produces design processes and artefacts *as well as* research that reflect upon the usefulness and quality of those design processes and products. Neither SE, nor HCI as described above fulfil this ambition, SE research being dominated by prescriptive research, and HCI by a more descriptive orientation, but we do believe, that being aware of the nature of the intra- as well as the interdisciplinary conflicts and differences in the fields is an important step towards resolving the differences. Adopting the idea of a design science may help solve interdisciplinary conflict in teams of software developers including HCI professionals.

Adopting the idea of a design science would, however, not resolve the issues that hide below the apparent similarity between lifecycle models from different disciplines. For example, what is the design outcome? Is it industrial art expression, is it a piece of software or

is it a change in an organizational process? Different answers suggest different process models. For example if you consider the design of new organizational work procedures, you would probably select a lifecycle development model that involved the workers, if nothing else as participants in training and as co-owners of the new procedures. The choice of lifecycle models for the design process influences what it is that is designed, because the choice underscores in different ways that software designers and end users all are constructors of the design artefact in all its aspects. It is this shared constructive aspect of all the eight cell positions in the design framework presented above that makes information science a design science in the sense of Simon.

Conclusion

It is a strong argument that a unified process which includes HCI and SE approaches is possible and needed. The SE processes have gone through a maturity curve over the past three decades and new approaches are constantly suggested. A 'truly' unified process integrating the activities of all disciplines contributing to designing software products is all set to emerge in the next few years. Such an integrated process cannot only improve the design quality of the products, it will also optimize on the effort required to build such products.

Currently, however, due to poor interdisciplinary communication and understanding, there is a gap between HCI and SE, which reflects a focus on design process versus artefact in use. The HCI/SE gap is furthermore complicated by the different understandings of science ranging from normative to descriptive that are prevalent as much within the disciplines as across the disciplines. Finally, there is a difference between software engineering having little to say about the artefact in use, and HCI being mostly about the artefact in use.

Therefore the constituent disciplines should address the HCI/SE gaps in all their complexity. The framework may help researchers as well as multi-disciplinary design teams to cope with this by giving them a way to address their own positions in relations to other possible positions towards design.

References

9241-11, I. I. S. (1998). Ergonomic requirements for office work with visual display terminals (vdts) - part 11: Guidance on usability (No. ISO 9241-11:1998(E)).

Bansler, J. P., & Bødker, K. (1993). A reappraisal of structured analysis: Design in an organizational context. ACM Transactions on Information Systems, 11 (2), 165-193.

Baskerville, R., & Pries-Heje, J. (2004). Short cycle time systems development. Information Systems Journal, 14(3), 237-254.

Battle, L. (2005). Patterns of integration: Bringing user centred desing into software development lifecycle. In A. Seffah, J. Gulliksen & M. Desmarais (Eds.), Human centred software engineering: Springer.

Beck, K., Beedle, M., Bennekum, A. v., Cockburn, A., Cunningham, W., Fowler, M., et al. (2001). Manifesto for agile software development. Retrieved September 14, 2006, from http://www.agilemanifesto.org/

Card, S., Moran, T. P., & Newell, A. (1983). The psychology of human-computer interaction: Hillsdale, NJ: LEA.
Clemmensen, T. (2006). Whatever happened to the psychology of human-computer interaction? A biography of the life of a psychological framework within a HCI journal. Information Technology and People, 19(2).
Clemmensen, T., & Nørbjerg, J. (2003). Separation in theory, coordination in practice - teaching hci and se. Software Process – Improvement and Practice, 8, 99-110.
Fitzgerald, B. (1996). Formalized systems development methodologies: A critical perspective. Information Systems Journal, 6, 3-23.
Friedman, A. L. (1989). Computer systems development. History, organization and implementation. Chichester: John Wiley and Sons.
Gillan, D. J., & Bias, R. G. (2001). Usability science. I: Foundations. International Journal of Human-Computer Interaction, 13(4), 351-372.
Guindon, R. (1990a). Designing the design process: Exploiting opportunistic thoughts. Human-Computer Interaction, 5, 305-344.
Guindon, R. (1990b). Knowledge exploited by experts during software systems design. International Journal of Man-Machine Studies, 33, 279-304.
Hartson, H. R. (1998). Human-computer interaction: Interdisciplinary roots and trends. Journal of Systems and Software, 43(2), 103-118.
Hevner, A. R., March, S. T., & Park, J. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75-105.
Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies, 64(2), 79.
Iivari, N. (2004). Enculturation of user involvement in software development organizations - an interpretive case study in the product development context, Proceedings of the third Nordic conference on Human-computer interaction. Tampere, Finland: ACM Press.
Jacobson, I., Booch, G., & Rumbaugh, J. (1999). The unified software development process. Boston: Addison-Wesley.
James, W. (1890). The principles of psychology.
John, B. E. (2003). Information processing and skilled behaviour. In J. Carroll (Ed.), Hci models, theories and frameworks - towards a multidisciplinary science (pp. 55-101): Morgan Kaufmann Publishers.
Jokela, T., Iivari, N., Matero, J., & Karukka, M. (2003). The standard of user-centered design and the standard definition of usability: Analyzing iso 13407 against iso 9241-11, Proceedings of the Latin American conference on Human-computer interaction. Rio de Janeiro, Brazil: ACM Press.
Kapor, M. (1991). A software design manifesto. Dr. Dobb's Journal.
Kelly, T. (2001a). Innovation – lessons in creativity from ideo: Currency Doubleday.
Kelly, T. (2001b). The ten faces of innovation: Currency Doubleday.
Madsen, S., Kautz, K., & Vidgen, R. (2006). A framework for understanding how a unique and local is development method emerges in practice. European Journal of Information Systems, 15, 225-238.
Mayhew, D. J. (1999). The usability engineering lifecycle: Morgan Kaufmann Publishers.
Mayhew, D. J., & Bias, R. G. (1994). Cost justifying usability: Morgan Kaufmann Publishers.
Naur, P. (1972). An experiment on program development. BIT, 12, 347-365.
Naur, P. (1985). Programming as theory building. Microprocessing and Microprogramming, 15, 253-261.
Naur, P., & Randell, B. (1969). Software engineering. Brussels: NATO Scientific Affairs Division.
Orlikowski, W. J., & Barley, S. R. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other. MIS Quarterly, 25(2), 145-165.
Preece, J., Rogers, Y., & Sharp, H. (2002). Interaction design: Beyond human-computer interaction: John Wiley & Sons.
Pressman, R. (2005). Software engineering – a practitioner's approach.
Roberts, P. H., Archer, B., & Baynes, K. (Eds.). (1992). Modelling: The language of designing (Vol. Design: Occasional Paper No.1.): © Loughborough University.
Royce, W. W. (1970). Managing the development of large software systems. In unknown (Ed.), Proceedings, ieee wescon (0 ed., pp. 1-9): IEEE TRW.
Simon, H. A. (1996). The sciences of the artificial (3 ed.). Cambridge, MA: MIT Press.
Stolterman, E. (1991). How system designers think about design and methods. Some reflections based on an interview study. Scandinavian Journal of Information Systems, 3, 137.

Walz, D., Elam, J. J., & Curtis, B. (1993). Inside a software design team: Knowledge acquisition, sharing, and integration. CACM, 36 (10), 63-77.
Wirth, N. (1971). Program development by stepwise refinement. Communications of the ACM, 14(4), 221-227.
Wundt, W. (1874). Principles of physiological psychology.
Yourdon, E. (1989). Modern structured analysis. Englewood Cliffs: Prentice-Hall.

INFORMATION AT A DISTANCE

LARS-ERIK JANLERT

Department of Computing Science
Umeå University
+46 90 7866131
lej@cs.umu.se

Abstract A systematic approach to the issue of the degree of availability of information requires that we abstract from physical circumstances in a way properly attuned to implementation issues. *Informational space* is taken to be the physical world measured and structured in terms of *informational distance*, which has seconds as unit of measurement and is oblivious to the manner in which information is "produced": local sensing, memory retrieval, computation on demand, or communication with remote information sources, or any combination of such methods. The *availability profile* for an agent is the spectrum of all informational distances, and may be identified with the agent's *informational state*. The location of an agent in informational space is suggested to be (over)determined by the availability profile, and the location of a piece of information (individuated by it being able to satisfy a particular request) to be determined by it being immediately available.

Introduction

It is common these days to talk of an information world, or even many, as an alternative or complement to the material world of mundane existence. Although *cyberspace,* in Benedikt's exposition 0, is said to parallel the physical world (with Popper's notion of World 1, 2 and 3 brought in to illuminate the relation 0), and by and large seems to be *about* or in other ways linked to the physical world—the popular message is that of physical independence, of freedom from physical fetters and actual,

physical circumstances, lending cyberspace its utopian (and sometimes dystopian) flavor. The world of ideas made visible and tangible, yet uncorrupted (and unrestrained) by matter. In summary, we may perhaps say that, *content-wise* the information world is *in principle* independent of the physical world (the freedom of thought, the power to fantasize) but *in practice* it is dependent on it, both because of a lack of other things to think about, and because of an understandable wish to make it do some useful work pertaining to this, our physical reality.

Yet, if there is an information world, it is in a concrete sense most certainly and in principle dependent of the physical world. Information, as far as is known, cannot exist without a physical carrier; it cannot be used and acted on without a physically realized agent; and it cannot be accessed without some physical process. Information always needs implementation. Although there is great freedom content-wise vis-à-vis the physical world, there are effects on content, on the access of content, from the implementation. We become concretely aware of these effects in our daily lives, e.g. when trying to quickly locate information that suddenly has become important to us. How do we reconcile the perfect world of ideas with the mundane real world? This is not just a theoretical but very much also a practical question.

Granted that the very point with different notions of information obviously is to abstract from physical realization 00, there is yet some choice as to *how* the abstraction is done. Can it be done in such a way that information and the “information world” come out as intelligibly related to the physical world? Compare e.g. the current confusion as to where information “is,” and the general awkwardness and uncertainty as to how to arrange information in the most useful and efficient way for our various purposes, which is becoming more and more urgent as more and more of our economy and personal well-being seems to depend on acting on the best available information in a situation where the available information is growing faster than we can think. The topic of this paper is the prospect of defining informational space in analogy with physical space in a way that puts the informational world in rapport with the physical world (without giving up power of abstraction). The main conceptual tools for this attempt are the graded availability of information, and the notion of an information availability profile, an information “spectrogram.”

To make this work, there is a need to abstract from physical circumstances in a way properly attuned to implementation issues. Whereas mathematics, logic and philosophy, in matters of information, generally can be said to be implementation innocent—going rather directly for the abstractions without wanting to tarry to consider physical information carriers in more detail—that is patently not the case with computing science, which abstracts from implementation and studies how computation may be implemented with equal dedication. It is the stance adopted here.

Informational space

If we understand **physical space** to be the physical world measured and structured in terms of physical distance 0, we may similarly understand **informational space** to be the physical world measured and structured in terms of informational distance. Such an approach means that we do not start with the assumption that there is an informational world apart from the physical world, but still strive to define an informational space distinct from physical space. Physical space and informational space can be viewed as two different abstractions of the same physical world, the “real” world. That is the general idea of the present proposal, and if successful it will guarantee at least a minimum of coherence between informational and physical space, being two abstractions of a common object.

Physical distance we can think of in terms of degree of physical presence. If an object is physically present, then the distance to it is zero. Objects that are at a distance are not present, not (literally) at hand, and the more distant the less present, the less at hand, the less accessible. Putting it this way makes it easier to introduce the concept of informational distance by analogy.

Viewing distance as inverse degree of presence, makes it natural and convenient to measure distance in terms of travel or transport times. There is a simple proportionality between time and distance given that speed is known: the distance to something or someplace can be viewed as the time it takes to make it present. (We may naively note that there are two basic ways of making something present: having it brought to us, or moving ourselves to it.) Physical space can be measured by rigid rods as well as by light particles traveling at a constant speed, and the latter is clearly the more practical alternative for large-scale charting of space.

Informational distance, then, is defined as the inverse degree of informational presence, being-at-handness or availability of some piece of information. Informational distance abstracts from procedures and implementation details. It is oblivious to the manner in which information is “produced” or “made present”: local sensing, memory retrieval, computation on demand, or communication with remote information sources, or any arbitrarily complex combination of these and other, similar methods. The inverse degree of presence of information is the time it takes to “produce” it, make it immediately “at hand,” and the natural unit of measurement is simply *seconds*. In other words, we abstract from everything in the implementation of information except the time taken to have it present. As long as we are not considering aspects of information such as value or validity—which are very deliberately left out of the picture in the current space project—this agrees well with common experience. Modern information technology often puts us in a situation where we don’t know exactly how the information is produced: whether it is computed just when we make a request for it, or whether it is retrieved from local memory storage, or perhaps fetched from some

remote central server. The procedure is also typically different at different occasions; for example, sometimes using cached information for what is basically computed information, or alternating between different servers (not to speak of network paths). In most cases, we do not care to know which way it is; we just want the information as quickly as possible.
In practice, of course, we will often have incomplete knowledge of informational distances, and uncertainty and misjudgments can become a major practical problem. If we have a particular way of getting a certain piece of information in *d* seconds, at least we know that *d* is an upper limit to the distance, but in many cases there could still be other ways of getting the information that are faster.

"Practical" space is impractical

That brings us to the question of practical versus ideal times for making something present. In the case of physical space one could conceivably imagine a *practical* physical "space," structured according to how long *in practice* it would take to move or transport objects between different locations. Practical distance would depend on local particularities as to accessibility, obstacles, and means of travel and transport; the shape of space would be irregular and change frequently when objects were moved or means of transportation were added or changed. This does not make for discovering regularities and making useful generalizations 0.
The standard concept of physical space, is of course an *ideal* space, which is more manageable both from a scientific and a pragmatic point of view. Physical space is by and large invariant and has a simple topology compared to "practical" space. It is in terms of this ideal space that the basic mechanical and kinematical laws of physics can be formulated. As we know, this idealized concept of space will also serve as a usable basis for analyzing practical world situations, although its power to do so may be somewhat overrated: the profusion of qualities and details of the physical world means that pure physics is not very practically usable in predicting the exact trajectory of a falling autumn leaf or exactly where a stick will break under pressure. At least we can conclude that the leaf *will* generally fall, and the stick *will* break, under specified circumstances.
The conclusion with regard to informational space seems obvious: although the particularities of an information world certainly are interesting, not least from a practical point of view, for a cleaner and scientifically as well as practically more negotiable approach in the long run, we should attempt to make the abstraction of an ideal information space.

Space and world

In conclusion, **space** is taken to be an abstraction from a **world**, where the particularities of objects and their placements are disregarded, i.e., we want the abstraction to be by and large independent of such

contingencies. The success of a proposed space abstraction, its legitimacy as a *space*, in other words, largely depends on the degree to which it achieves that goal, as discussed above. Aggregated effects of world on space may still be covered and considered legitimate properties of the space. Examples from physical space theory include Mach's revealing comment on Newton's bucket experiment, which Newton used to argue for the absolute space of classical physics, in particular a favored reference system with regard to rotation and acceleration. Mach pointed out that inertia might be explained as an effect of the great masses of the actual world: inertia of an object arises in its relation to the average motion of the total mass of the universe, the aggregation of all objects. A more recent example is the physical space of general relativity, which is at least popularly understood to have the property that mass can bend it; globally such that the size, curvature and topology of physical space depend on the amount and overall distribution of matter (objects), and even locally, with sometimes spectacular effects such as those in the vicinity of a black hole.

In cyberspace "theory," Benedikt makes the analogy by discussing the "density" of information in cyberspace, and how high information density should have the effect of slowing down movements 0. The *space* in his version of cyberspace, is a human artifact in the sense of a construction (and not just in the sense of a deliberately chosen abstraction) and should preferably be designed, he suggests, so that movements in it *incurs* a cost proportional to distance.

Space can be understood as a container, as an absolute space, as Newton did, or as a measuring system (relative space). The approach taken here is to see space as a way of accounting for a world in terms of relative distance relations, which allows us to exploit the analogy between physical and informational space.

Computation bends informational space

In a world without computation (only communication), informational space would in a sense parallel physical space: informational distance would be essentially proportional to physical distance. To be sure, the situation is complicated by the fact that different implementations and carriers of the same information are, per definition, identical *qua* information, whereas (normal) physical objects are not considered identical even if they happen to be very much alike or even indistinguishable. That makes many "information objects" take on a fragmented and scattered shape from the physical space point of view. Information may be recorded, duplicated and communicated, i.e. transported, which means that e.g. the distance to "the temperature of *A*" is the distance to whatever instance of information regarding the temperature of *A* happens to be closest by: it could be a thermometer at *A*, it could be a note about the temperature on a piece of paper at some other location, it could be a remote display

coupled to the thermometer at some other location, etc. Still, the informational distance, the access time, is correlated with the physical distance to the appropriate physical carrier.

With computation, the topology of informational space will deviate in more interesting ways from the topology of physical space. Instead of fetching information from afar, there is the alternative of computing it from information closer by (and some information may even be available only through computation). This is what makes informational space topology quite different from that of physical space. At this point, one might reasonably raise the question whether not simply taking part of information always involves something similar to computation on the part of the agent accessing it—I will get back to this.

As an aside, we might attempt to work the analogy in the other direction and consider the idea of "derived" physical objects. Would that undermine the claim that informational space is topologically quite different from physical space? It would, were it not for two things. First, in all but a few special cases the procedure to manufacture an object from parts or ingredients is obviously so much *slower* than getting a ready-made, even from afar. Second, there is the richness of physical objects already mentioned, which means that no two (macro-sized) objects are exactly alike, exactly interchangeable; not at all like information, where the very idea is that of information carriers totally interchangeable with regard to content—and content is all that ultimately matters, once we have the information in our hand.

Comment [Ll
could make
if equal ph
objects wer
identical:
production
from availa
materials,
with approp
procedures,
chemical or

General facts considered as computational formulas

In order to compute, there needs to be some rule, some formula for the computation. Any report of a general fact, such as that lead melts at 327.46 ˚C, can be viewed as a formula for computation, a program. E.g., to compute whether some piece of lead *A* is melting, using available information about the temperature of *A*: if the temperature is at least 327.46 ˚C, it is melting; if not, it isn't. Hence, if the informational distance to the temperature is short, and the piece of lead far away, the informational distance to the state of aggregation of the piece of lead may still be small, given that the informational distance to the fact about the melting point of lead is small.

This goes a way towards putting informational and computational resources on an equal footing, emphasizing the timing aspect. Still, deconstructing computation into components of information is expected to always leave some active component; processors, computational agents of some sort.

A piece of information

An abstraction of information in the style of propositions (e.g. infons 0) has some well-known nice properties like combinability, compositionality,

etc. One serious disadvantage with such approaches is indeterminate availability: propositional content does not individuate information in a way useful for measuring up informational space. That a physical information vehicle V enables us to meet a certain information request in time T by delivering a certain propositional content P, does generally not imply that V can meet a different information request answerable by the same propositional content P in the same time T. E.g. if P is 'Paris is the capital of France,' in most implementations V, the request for the capital of France would take different time to satisfy by V than the request for the country that has Paris as its capital.
As an alternative I propose that a particular **piece of information** is to be individuated by it being able to answer or satisfy a particular request—so that information *needs* rather than supplies determine the units of information. (This reversal may also help to take some of the edge off the apparent but spurious self-sufficiency of information carriers.) The **location in informational space** of such a piece of information is determined by it being immediately present (null distance) at this location. There is little choice in defining the location of information in any other way, given the present setup, but one not so desirable consequence will be that a single piece of information can appear scattered in informational space (not just in physical space, which is already taken for granted, see above). That, however, is a price we may have to pay to keep informational space in touch with physical space to the degree that informational space can be used for practical applications.

Informational state

The **availability profile** for an agent is defined to be the complete spectrum of all informational distances (for all information). I propose to identify the availability profile with the agent's current **informational state**: the availability profile gives a full account of the degree of availability of any information, including information that is very close, as well as information that is very remote and possibly does not really play a role in the agent's current activities. In other words, it covers all there is to say about the agent's current state with regard to information. Does it say too much?
The distinction between internal and external information has faded away in this definition, so that external information is as much a part of the informational state as internal information. This is as it should be: it agrees with our common experience (you get wiser in a library), it reflects the distributed way in which modern information technology works, and, as it happens, it fits in nicely with current theories of cognition such as distributed cognition 0 and situated action 0. Also, we already know from experience that external information can be informationally closer than internal (not considering now the frequent cases where the agent is not informationally self-sufficient with regard to the requested information);

e.g. instead of trying to recall a name or a fact, we can save time by looking it up, given that we have the right facilities available; instead of adding a list of numbers by mental arithmetic, we find it quicker to use a pocket calculator. From a pragmatic point of view, when situated in a particular environment, the boundary between internal and external becomes irrelevant.

Agent anatomy

The question of the agent's own—and now "own" does not necessarily mean "internal" in the physical sense above, we are talking about the *informational* boundaries—informational and computational resources arises for another reason: having defined informational distance in terms of the time it takes to make the information present, we have implicitly introduced an agent, and consequently need to say something about its capabilities. A book in Chinese about feng shui may quickly inform a reader about the proper place to put the kitchen table (according to feng shui, that is)—if the reader has adequate eyesight and can read Chinese. If not, the requested information will not be very present.

From the point of view of measuring and charting informational space, very thin agents, some kind of standardized probe with minimal informational and computational resources, is ideal. It will be somewhat like using photons or electrons to explore physical space. When we want to move a little bit further towards practical applications we will need to consider a wider variety of agents, some equipped with extensive informational and computational resources of their own. Still, it may be possible to strip a "thick" and resourceful agent down, layer by layer, each layer holding resources external to the inner layers and extending their capabilities. It may happen, though, that such an operation will not converge on a common core but diverge into clusters or societies of smaller agents. From this point of view there seems to be no single way of drawing a line between agent and world; rather there are many ways of drawing the line, allotting less or more to the agent and to the world, respectively, as we find convenient. The agent passes gradually into the world, mirroring the gradual passing of present information into distant information.

Relation to physical moves

Given agents that can move in physical space, there will be differentiation between information that is moved along with the agent, i.e. remaining close in informational space, and information that is not. That may suggest a certain physical boundary between agent and world, if we did not have one before. Yet, a software agent might be considered to move between different host computers, located at different points in physical space, clearly without really moving any physical parts. Considered in its

nomadic existence, however, we may find it convenient to distinguish the informational boundary outlined by its physical relocations.

Movement in physical space will generally affect the informational state of the moving agent: some informational resources will come physically closer, and thus informationally closer; some will become more remote.

Changes in the physical environment will generally affect informational state, assuming that we are primarily interested in agents of limited capabilities involved with the physical world: sentient beings, informationally connected entities. There should be room in informational space also for eremite agents informationally cut off from the rest of the world, and possibly for omniscient, omnipresent beings, too, but they are presumably of marginal interest other than as a way of testing extreme points of the theory.

Given agents that can make some kind of change in their physical environment—which includes any informational change, since informational change is impossible without physical change—they will consequently potentially affect the informational state of other agents. Simply moving in physical space can have an effect on their informational state.

Location overdetermined by informational state

The working hypothesis is that **agent location** in informational space is determined – and overdetermined—by the availability profile. Given a certain minimal richness of content of the world, different physical locations imply different informational states (following the world-involvement principle), whereas different availability profiles may be compatible with the same location if we have agents of some nonnegligible thickness. As the agent moves in physical space its informational state will change, generally entailing a movement in informational space as well. Self-movement generally implies a massive relocation of information sources relative to the agent, as we can expect most other objects to change their position relative to the agent. On the other hand, we may expect most of those objects to be quite some distance away so that the relative change is small and they affect the informational state only marginally. Objects that are close will give rise to larger changes, but we may expect those objects to be few, relatively speaking, so the overall effect on the informational state may still be marginal (yet, that change, small in the larger perspective, could be very important in a particular situation).

An agent may of course move in informational space without moving in physical space, by the accruing and rearranging information near at hand, by improved communications, added computational resources, etc; neither of which, however, is possible without physical change of the agent and/or its environment. Note that some changes are minor, in the sense that they affect the availability profile only marginally (and so has

little affect on the position in informational space), whereas some changes are major: they affect considerable areas of the availability profile; computational changes will typically have such an effect. Everyday examples would be to learn a new language, or learn to do arithmetic calculations.
Just as it would not be a good idea to define physical space so that a specific physical resource such as 'gas station' is at a single point in space (assuming, of course, that there are more than one instance of gas stations in the world), from the above discussion it does not appear likely to be a good idea to insist that a certain informational resource is at a *single* point in informational space. For reasons of simplicity, we seem forced to accept the at least initially rather unattractive alternative of distributed objects. The tendency of abstractions to unify and scatter at the same time is manifest also in the propositional approach to information, and perhaps we must accept that a "piece" of information is much more like a property than a thing.

Conclusion

I have presented a sketch, an outline of an informational space theory. It is clear at this point that the difficulties are many, and it will be no easy task to work out the details so as to reach the lofty goals of providing an abstraction for the informational view of the world comparable and compatible with what the concept of physical space does for the physical view of the world. In the meantime, I believe that the set of concepts I have introduced—in particular informational distance, availability profile, and the view of informational space as an alternative abstraction of the physical world—have an interest and a value of their own, and may well be worth developing for what they may do to improve informational thinking and reconcile it with physical thinking, as well as for their potential to build a full-blown theory of informational space.

References

Barwise J, Perry J. Situations and attitudes. Cambridge, Massachusetts: The MIT Press; 1983.

Benedikt M. Cyberspace: Some proposals. In: Benedikt M, editor. Cyberspace: First steps. Cambridge, Massachusetts: The MIT Press; 1991.

Floridi L. Information. In: Floridi L, editor. The Blackwell guide to the philosophy of computing and information. Oxford: Blackwell; 2004.

Hempel C G. Philosophy of natural science. Englewood Cliffs, New Jersey: Prentice-Hall; 1966.

Hutchins E. Cognition in the wild. Cambridge, Massachusetts: The MIT Press; 1995.

Machlup F, Mansfield U, editors. The study of information: interdisciplinary messages. New York: Wiley; 1983.

Popper, K R. Objective knowledge. An evolutionary approach. Oxford: Oxford University Press; 1972.

Reichenbach H. The philosophy of space and time. New York: Dover Publications;1958.

Suchman, L A. Plans and situated actions: the problem of human-machine communication. Cambridge: Cambridge University Press; 1987.

HOW MAY ONE OVERCOME THE ARTIFICE OF *KNOWLEDGE*, WITHOUT ENDANGERING COMPUTER SCIENTISTS?

DIGITAL DOCUMENTS, KNOWLEDGE *AND* COLLECTION.

Francis Rousseaux
Laboratoire CréSTIC (EA 3804)
Université de Reims URCA
francis.rousseaux@univ-reims.fr

Thomas Bouaziz
ExperiensS
thomas.bouaziz@experienss.org

Burkhard Mayer
Ecole Centrale d'Electronique
mayer@ece.fr

ABSTRACT: *When in 1982 Allen Newell invented a new definition for knowledge, thus enabling computer scientists to reunite the two founding sides of artificial intelligence, he was probably far from imagining the huge success his proposition would come to. The manner in which both digital documents and the interactive tools allowing access to their content were assumed was to be seen in a different way, hence opening up to new innovative applications. However, analysing systems aimed to help in document interpretation over the last fifteen years and based on Newell's hypothesis, has lead to something quite unexpected: to conceive most of these systems, engineers have in fact been trying to* deconstruct *the hypothesis. Typically, the notion of* collections *has been preferred to that of* knowledge. *Yet perhaps the notion of collection could only appear once that of knowledge had been taken into account, distorted and finally subverted.*

KEYWORDS: knowledge, collections, digital data mining, machine learning, music, maps.

1. Introduction

The conception we have nowadays results from a tradition that considers documents were knowledge containers. This figure of containment yet widely exceeds the scale of the phenomenon it is meant to describe.

Why don't we try to remain closer to the very phenomenon, even if this stands for refusing *ad hoc* explanations devoid of any stimulating effect? Here is what we experience first hand: when successfully carried out, the process of "getting to know" a document induces liveliness and an animation of thoughts. This phenomenon thus initiates a desire to "know more"; it brings forth a determination to be confronted with the document (or to gradually shift to other documents) and, when document reception results from the figures of attraction (continuance/repetition) rather than repulsion, it eventually results in the production of new documents.

Computer scientists are often fervid yet innocent advocates of this tradition which firmly argues that documents are knowledge containers. It's not that they have been hired by some activist promoting such statement, nor are they particularly interested in this debate. It rather seems that the very history of computer science, originating at the same time as that of Artificial Intelligence (AI), logically leads them to holding a tacit position on one definitely strategic ground.

This paper aims to explore the invention of *Knowledge*[67] in the field of computer science; we should argue that it stands for the origin of the *biased vision* most computer scientists have on digital issues.

2. The invention of *Knowledge* in the field of computer science

Despite its yet short existence, computer science nonetheless originates from a rich and complex history; it began with the Cold War, and was, at the time, a vast and ambitious transdisciplinary project with an extremely meaningful name: *Artificial Intelligence*, thus highlighting the second meaning *intelligence* has in the English language.

The research field has been widely influenced by Alan Turing's founding works and has gained strength with Herbert Simon's ecstatic prophecies, not to mention many other significant contributions. Everyone knows that. Yet the paramount part Allen Newell played is often forgotten; building *Knowledge* from a computer science point of view, he literally

67 In order to emphasise the fact there is no concurrence between the notion of Knowledge computer scientists understand and the commonly assumed notion of knowledge – despite the use of the very same word by computer scientists aiming at forcing such concurrence – a capital K letter should be adopted so as to highlight the intrusive meaning of the word.

invented a new meaning for a prevalent notion in metaphysics. It meant building an operating and favourable notion for designers and programmers of computerised systems while trying to convince them they were holding the sacred grail metaphysics had been trying to define since the dawn of time. Many discerning computer scientists have tried setting up innovations grounded on Newell's assumptions, many more have been influenced by his ideas, while totally unaware of these origins.

2.1 The state of artificial Intelligence in 1982

What was the state of AI when Newell began writing his famous *"Knowledge Level"* (Newell, 1982)? To say the least, the AI research programme was on the verge of collapsing, torn as it was between two sides. Its field of action was a utopia ironically established by Alan Turing via the two figures he had set: his Machine and his Test (Turing 1995).

A Turing Machine is a virtual logical machine, later designed architecturally by von Neumann (von Neumann, 1996), and eventually materialised in the silicon of computers. It allows the operationalisation and simulation of some temporal and/or causal phenomena via the assimilation of the necessary reason *(Modus Ponens* or deductive reasoning) with causality, then via the automatic effectuation of the logical inference transformed into calculus (Turing, 1939). In a Turing Machine, AI lies in a corpus of programming techniques that have been specified to tackle issues of Problem Solving[68], the ones Newell engaged in with his *General Problem Solver* (GPS).

As for the Turing Test, it puts together the intersubjective dialogue and the mystery contained in its *continuation*: an interlocutor is deemed intelligent as soon as he answers back. This is how a Human may personify an artificial interlocutor, as soon as the latter is capable of continuing the conversation in time and of remaining at the mercy of the human speaker. Hence, as far as the Turing Test is concerned, AI is a phenomenological investigation on the issue of the dialoguing subject.

Newell totally refused the reduction of Turing's ambitious research programme to a type of engineering meant to serve the Theory of information. As a result to his own culture as an engineer fascinated by the engineering side of AI, Newell got involved with the Machine, suggesting computers be considered as layered systems *(Symbol Level)*; he also suggested adding a superior layer (the *Knowledge Level)* built so as to reach the Test side of AI and to finally get rid of the menacing crisis. This also meant putting a stop to the seeping questions that were gradually cutting into the area.

68 In this paper, all words referring to specifically classified notions in computer science are to be written with a capital letter.

2.2. *The Knowledge Level*: Allen Newell's proposition

Newell invented *Knowledge* in order to bring a solution to the Human–Machine problematic issue found in artificial Intelligence. His answer to the controversial question "Who, between Man and the Machine, is intelligent?" is "Let them both become intelligent together; as a multiagent hybrid couple/group/organisation, Knowledge being the junction enabling the coupling and the interdependency between the human and the machine".

In computer science, Knowledge refers to the condition of possibility for the hypothesis of Newell's Knowledge Level. An interactive Human-Machine cooperation is based on a principle of rationality (I like expressing this principle the Montaigne way: tell me what want, what can, what know, I'll tell you what do). In that it may be handled by Man – who can therefore assume his thinking as a rational and finalised tool – Knowledge exhausts/reduces/describes the phenomenon of thought(s). In that it can be represented and implemented in computer science systems, Knowledge gives computers information on the situations and levels of freedom of human actions, the machines may therefore mobilise the operation of instantiation as well as logical inferences in order to embark on various types of rational reasoning.

Newell's talent was precisely to go through with the conception of this rescuing move: naming the area dedicated to the dual monster[69] – whose den had already been localised by Turing as being the Human-Machine interaction.

2.3. Consequences of Newell's move: feedbacks

Newell's Knowledge is logical – teleological, to be more specific – *out of time and out of human desire*. His Knowledge may be regional/domanial, or job oriented, yet it cannot be located since it is literally uninhabited. Newell's Knowledge claims the reduction of the trivial knowledge extended in time – both *narrative* and *discursive* – of the human mind, in order to categorically stifle it with an instantaneously finalised rationality.

When taken at face value, Newell's proposal surely impoverishes the human thought, yet it also displays productivity and a capacity for innovation of its own. What's more, ways to deconstruct it may be

69 Like the Roman God Janus, Knowledge has two faces, one turned towards the Symbol Level of the Machines, the other turned towards human actions which, according to Newell, are always rational and finalised.

explored, for instance focusing on less severely reducing inventions such as the notion of *collections* (Vignaux, 2004) – a more precise singular notion, happening here and now, in a field of dynamic attraction. The course may thus become choreographic/scenographic and hence relinquish the topological characteristic – assuming it may still be mapped – of its own inscription, as we shall later discuss.

Analysing examples of computerised systems we are familiar with – having taken a more or less important part in their conception/realisation – will lead us to a critical study focusing on the productivity of Newell's Knowledge. The analysis should be based on browsing systems found in digitalised collections of music extracts – namely LE MUSICOLOGUE and CUIDADO – and should also focus on situation control systems – CHEOPS and VIRTUALIS in particular.

3. Browsing digital musical documents

Setting up a browsing system via digitalised musical documents entails preliminary complicated problems, mainly in terms of acquisition and restitution but also relating to representation and to the Human-Machine interface. When such problems are finally overcome, the major difficulty arises: making use of tools based on Newell's *Knowledge Level* in order to subvert his initial propositions, and lessening a *priori* instantiation and classification so as to reach situation similarity and a collection encompassing *singularities*.

3.1. Introducing LE MUSICOLOGUE, a music browsing system.

The MUSICOLOGUE system was conceived and realised by a small team of computer scientists and musicologists between 1987 and 1990. Among the various ambitions of such a system, one of them was that the system should suggest a new piece to work on (Rousseaux & Saoudi, 1991) to a student who had just practised music dictation on a certain piece. This being performed with optimal coherence in terms of corpus, namely a collection of elaborate exercises adapted to the student's improvement.

The subsystem in charge of practically suggesting new pieces to work on – a process depending on the piece currently being dealt with and the student's own difficulties – incited us to use the DISCPLE system. The latter had been developed a few years before in the automatic Learning research team of the Paris 11 University – a project we were also involved in. (Kodratoff, Tecuci et Rousseaux, 1987).

DISCIPLE is a learning apprentice help system for browsing in a logical problem solving process which evolves via goal regression, and

mainly used in planning. DISCIPLE learns by searching how to put together practical Knowledge, namely rules for decomposing problems, and theoretical Knowledge, represented in a large semantic Network (Brachman, 1979).

DISCIPLE had been developed in a theoretical formalist logic of learning, hardly taking into account the Human-Machine interaction, and reducing it to a vote-catching sort, typical of Expert Systems: the Human is expecting solutions; some are suggested by the system; yet only when the computerised system fails to provide some, is an expert sent to engage in an updating and learning Knowledge process guided by the machine.

3.2. Seeing the corpus as a collection of pieces worked on

LE MUSICOLOGUE helps the student build a collection of pieces he worked on. Collecting is more original a word – as relating to an origin or beginning – than categorising. It goes with time, a sort of *Lebensvelt*. It is mostly true when working on music pieces, as in this very case the key to success is the continuance of an activity which never stops nor repeats its object. Indeed, it expands into sequences of objects standing for the navigation path of a collection (Rousseaux 2004), quite similar to when one builds a collection of art works (although the appropriation of temporal objects cannot be compared to the appropriation of spatial objects). Yet, if the impression an activity leaves behind in the world is nothing but its continuance, how may one set up a Human-Machine dialogue? Which medial Knowledge should it be built on?

In the insertion environment of LE MUSICOLOGUE, the student leaves impressions of his exercises – impressions or trails different from the preliminary selection of the piece he is working on –: both the evaluation of his work and his level in the corpus have been carefully thought of in order for the learning apprentice system to have grounds on which stimulating the student's interest by offering him a number of *interesting* pieces to work on, which the student may choose to select. Yet what about the act of purely listening to music – implying neither note taking no any other trace but the sole desire for its continuance –? Could a system which would offer the listener some help to set up a navigation path/collection be considered even if no goal outside the actual activity may be assigned to the system? This is precisely the objective set for the *Music Browser* Sony-CSL developed as part of CUIDADO, a European project coordinated by the Ircam between 2000 and 2003 (Vinet, Herrera et Pachet, 2002).

3.3. A study of the CUIDADO browsing system.

Music browsing inside large corpora of digitalised pieces is widely influenced by the notion of *genre* which itself originates from the need to physically choose the CDs one wishes to get in the various shelves and departments specialised shops contain. The end of the CD as a medium entails the end of the hegemony this purchase activity has had so far. It has also given rise to a number of competitive activities laying claims to the cut of indexation, hence the advent of plethoric and competitive index-linking regimes.

For this reason, CUIDADO's *Music Browser* presents – jointly with an editorial metadata indexation – the user with *cultural* and *acoustic* search possibilities. It also leaves aside the imposing of exclusive categorisation based on those types of index, while encouraging the user to shift via a search for similarities as transversal and interactive as he wishes (Pachet 2000 et 2003). The idea of collection is once more brought into play; the system offers the collector/listener opportunities combined on different levels and yet which may always be simultaneously activated if need be. He is therefore free to choose which he may locally have power on.

3.4. Partial conclusion and first teachings

The differences between the two systems – LE MUSICOLOGUE and CUIDADO – are not so much technical as epistemological. From one system to the other, there is a shift from (LE MUSICOLOGUE) a world filled with formal categories, in which the Machine tends to be in charge of the loop of interactive events, to (CUIDADO) a situation dealing with singular collections and in which the Human tends to remain responsible for – and the ultimate master of – the loop of events and the results of the Human-Machine system. In CUIDADO we do not even refer to results anymore but to *paths*, knowledge being here constantly engaged in a sustainable narration devoid of the need for exogenous or endogenous goals of the system, simply by shifting from/to similarities.

In LE MUSICOLOGUE the very ruled and utilitarian characteristic of the information system is precisely what enables it to come along in terms of instantiation, managing the content as facts instantiating its knowledge. Yet if what is aimed is an open system undefined by a primary use, which role should it play in the context? It is therefore necessary to tackle the issue differently as it is not possible to reduce contexts of uses to predefined generic cases. It is therefore necessary to shift from the particular to the singular.

4. Browsing through scenographic digital documents

In this second phase of system analysis, we should focus on achievements dealing, this time, with documents of a scenographic type.

Again by removing the explicit or implicit teleological requirements from the conception of the information system, we should introduce VIRTUALIS (2005) and highlight how this decision making help system explores situation shifting throughout the Human-Machine interactions.

4.1. The VIRTUALIS system: generating collections of interactions

VIRTUALIS is a system grounded on the idea that a performance may be considered as a collection/procession of interactions that are under constraint; setting up processes that retain several interactive exchanges may open the work – understood here as in Umberto Eco's open work – by densifying the interactive space (Rousseaux et Bonardi, 2004).

For instance, Alain Bonardi, the main conceptor of VIRTUALIS (Bonardi et Rousseaux, 2001; 2004), has set the system up in a play by Geneviève de Gaulle. The play staged a narrator delivering the whole text of the play – Valérie Le Louédec – and a dancer achieving a certain type of gesture directly inspired from the Noh theatre – Magali Bruneau – as well as a gigantic screen at the back of the stage on which mobiles were sketched; the mobiles were directly animated with the particular emotions felt in the narrator's voice. The immediate influence of the voice was thus jointly meditated via the screen; it therefore reached further remanence and a wider range in terms of temporal density. The screen induces reactions from both players, more particularly that of the dancer who adapts her own gestures to the movements and qualities of the image.

What we thought was interesting in such a project was to try to lessen instantiation (Rousseaux et Bonardi, 2004) – the very unthought-of concept in computer science – by suggesting the direction of the play should be conducted and specified by a shift in situation controlled by the situation itself, rather than by variations of instantiations in the ontology of characters and actions.

With an interactive data mining approach we may see the example as a specialisation of overall cases; other similar specialisations may be searched for yet without the help of a predefined ontology. The user therefore accepts shaping it in an *ad hoc* manner, using the machine's interactive help.

4.2. New conclusions and additional teachings

Through VIRTUALIS, Alain Bonardi and I discovered that technology escaped from the concepts it originated from, or to be more accurate, it deconstructed them – as Derrida defined it. Indeed, even immediately

after it surreptitiously started encouraging organisations in its work and conception methods, Newell's molecular and mapped knowledge could start being deconstructed. The notion clears the way for as much scenography – and as many choreographies – as experiences of the work; this development recalls Simondon's definition of *concretisation* (Simondon 1989); the innovation consists in deconstructing Knowledge and be done with its artifice, while still using the tools the dogmatic notion helped set up.

It was then becoming clearer that Newell's attempt was a fiction meant to work in the representation field, in order to freshly address the never-ending crisis of representation, in partnership with computers. The latter may potentially contribute to the elaboration of less frustrating representations for they encourage more sensorial and conceptual investigations. Computers should therefore be granted means enabling them to be part of the interactive mediation of representations.

Even with tremendously high rational requirements, it was probably very clever to immediately propose a radical solution: a two-sided type of Knowledge featuring a static side turned towards computers and a dynamic one turned towards humans. Yet today it is clear that a part of the dynamic one may be turned towards computers and that the hypothesis of rationality may even be alleviated – and should be – as long as one renounces the two-sided Knowledge to prefer a horizon filled with synthesis and desire, better inspired by the notion of Collection – as an art collector may experience it – than by the notion of rational Knowledge.

After Information, and after the form/substance Relation, it is now time to question Knowledge and Contents of digital documents.

5. Deconstructing *Knowledge* and the invention of *Collection*

By revisiting prior experiences in conceptions of intelligent systems helping with the interpretation of digital documents, we have noticed that the proposition for a description of knowledge in principle – outside any experienced situation – had left us often quite unsatisfied and that we had preferred that of shifting to and from situations, hence creating a Human-Machine interaction in time – through narrative forms – and thus enabling the setting up of motivated collections with lively and dynamic intentions.

Digital documents do not *hold/contain* knowledge. Their "acquiring knowledge" is much rather a process elaborating collections, targeting both its completion and continuance, both prospects being necessary together because they maintain their reciprocal possibilities.

Let us think about art work collections and about Gérard Wajcman's analysis (*Catalogue de l'exposition inaugurale de la Maison rouge,* page 89) on the status of excess in collections: "Excess in a collection does not mean disorganised accumulation. There is a founding principle: for a collection to be so – even in the eyes of the collector – the number of works needs to exceed the material capacities of displaying and stocking the entire collection at home. Someone living in a studio apartment may very well have a collection: he will only need to not be able to display at least one work in his apartment. It is for this reason that the reserve is one full part of collections. Excess can also apply to memorising abilities: for a collection to be so, the collector should be incapable of remembering all the pieces he possesses (…). In fact, he either needs to have enough pieces to reach the "too many" and to "forget" he had this or that one, or needs to be compelled to leave some outside his place. To put it in a nutshell, what makes a collection is that the collector should not have total power over his collection".

"A private collector's scene is not his apartment but the whole world. It's important to stress that the major part of his collection in not to be found at his place, his collection is yet to come, still scattered all over the world. Any gallery or fair represents the possibility of chancing on his collection yet to come." (Wajcman, *Collection*, p. 29) Also: "No one can ever look at "one collection" since it is not a whole work but an infinite series of singular objects, a piece + a piece + a piece, etc." (Wajcman, *Collection*, p. 28)

The process of extending a collection is potentially infinite even if the collection is necessarily undetermined, *temporarily* finished. Practically speaking, a collection ceases to exist as something else than a commonplace correlate whenever the collector loses interest in its extension: he then stops reiterating the acquiring gesture and/or the reconstitution of the collection in an intimate dwelling comes to an end. Both acts have the same essence: in order to keep the collection in an intimate sphere, the collector pays a visit to his sheep[70] and re-generates the collection, working on his very logic of growth, yet unaware of it. Re-production balances the collection's heavy trends and facilitates new links among the pieces, hence setting up new similarities that will eventually influence the acquiring logic. Strangely enough, desire becomes knotted

70 At the beginning of André Gide's Symphonie pastorale, the good shepherd who has welcomed Gertrude tries to dispel his wife's premonitory worries. He defends his peculiar interest in the young blind girl by spiritually recalling the most particular devotion implied in a secluded life of infirmity. When later in the novel the wife is surprised the shepherd abandons his own children, he hides his consciousness behind Matthew's Gospel and answers back that "each sheep of the herd, taken on its own, is more important in the eyes of the shepherd than the overall herd taken as a whole." I've always seen this as another collection metaphor; the shepherd sees the overall herd as an abstraction. As soon as action is needed for an endangered sheep, the figure of the herd fades away and gives way to the singularity of the needy sheep.

to difference. Object enter the collection via the *being different* predicate; they only become similar later on, as being different is what they have in common, hence setting up what Jean-Claude Milner calls a paradoxical class.

If after Simondon we may talk about material realisations, what about the genesis of symbolic systems and outgoing technical tracks originating from this? Could we possibly have practice paths and conception tracks? Shouldn't we begin considering a new technicality in computer science?

6. Bibliography

Bonardi A. et Rousseaux F. (2001) « Interagir avec un contenu opératique : le projet d'opéra virtuel interactif Virtualis ». *Revue d'Interaction Homme-Machine*, numéro spécial *Interaction et Documents*, Volume 2, n°1, juin 2001.

Bonardi A. et Rousseaux F. (2004) « A la croisée de l'informatique et du théâtre : similarité en intension *vs* en extension ». *Conférence Internationale ARCo*, Compiègne, décembre 2004.

Brachman, R-J. (1979) « On the Epistemological Status of Semantic Networks ». *Associative Networks,* Academic Press, New-York.

de Galbert, A. (2004) « L'intime — le collectionneur derrière la porte ». *Catalogue de l'exposition inaugurale de la Maison rouge,* Fage.

Kodratoff, Y., Tecuci, G. et Rousseaux, F. (1987) « DISCIPLE: a LAS for Weak Theory Domains. *Cognitiva »*, Paris, mai 1987.

Newell, A. (1982) « The knowledge Level ». *Artificial Intelligence*. Vol. 18, p.87-127.

Pachet, F. (2000) « A Taxonomy of Musical Genres ». RIAO, Paris.

Pachet, F. (2003) « Nom de fichiers : Le nom ». *Actes du séminaire STP,* MSH Paris.

Rousseaux, F. (2004) « Collecting or classifying? — Intimacy clarified: a collector's ritual ». Special Interest Group on "Philosophy and Informatics", *Professional Knowledge Management International Conference*, Ulm, octobre 2004.

Rousseaux, F. et Bonardi, A. (2004) « Music-ripping : des pratiques qui provoquent la musicologie ». *MUSICAE SCIENTIAE*, numéro spécial 2003-2004.

Rousseaux, F. et Saoudi, K. (1991) « Informatique musicale, pédagogie et communication ». *Revue de pédagogie musicale MARSYAS de l'Institut de Pédagogie Musicale*, dossier n°20 "Informatique et pédagogie", décembre 1991.

Simondon, G. (1989) « Du mode d'existence des objets techniques ». Aubier.

Turing, A. (1939) « Systems of Logic based on Ordinals ». *Proceedings of the London Mathematical Society* n°45.

Turing, A. (1995) « Computing Machinery and Intelligence ». *Mind LIX*, n°236, 1950; réédité dans les Collected Works of A-M. Turing, vol 3. *Mechanical Intelligence*; traduction française dans A-M. Turing, J.-Y. Girard, « La machine de Turing », Seuil, Paris.

Vignaux, G. (2004) « La notion de collection : genèse, développements, valorisations ». *Rapport interne CNRS du RTP 33 "Document et contenu : création, indexation, navigation"*.

Vinet, H., Herrera, P. et Pachet, F. (2002) « The Cuidado Project ». *Actes des conférences ISMIR*, Paris.

Von Neumann, J. (1996) « Theory of Self-Reproducing Automata ». London and Urbana, University of Illinois Press; traduction française : *Théorie générale et logique des automates*, Paris, Champ Vallon.

Wajcman, G. (1999) « Collection ». Nous.

INTERACTIVE COMPUTING AND CAUSALITY

Anne NICOLLE
GREYC UMR CNRS 6072
University of Caen
Tel : 33 2 31 56 74 34
Mail : anne.nicolle@info.unicaen.fr

Abstract : Interaction of computers systems with their users and their environment is the most important dimension of modern software. Peter Wegner proposed interactive foundations of computing and has developed them since ten years. It is a good background for empirical computer science and particularly to explain the behavior of software agents. At the opposite of functions who have to compute a result in a finite time, the run of almost interactive process is indefinite in time. What are the causes that explains how these processes product their effects ? All computational process has the same material cause: networks, memory and processors. The efficient cause is the function, either the interactive behavior, or the combination of them as they are executed on a processor. Their formal cause was the program who is public and may be studied exhaustively. But some processes called agents have an autonomous behavior which can't be explained without a final cause.

Introduction

In order to design computer software, two views are used: a structural view which describe objects and relations between them and a behavioral view. The paradigm of theoretical machines to model computation, like Turing machine, proposes a static study of the behavior: complexity and termination. Turing machines modelize early states of computer software. They are extended by persistent Turing machines in order to model the use of files and data bases. Interaction machines models modern computer software with graphical user interfaces and communication by networks. Part 1 presents this static view of behavior.
Then, part 2 presents a dynamic view of the behavior of computing

software, in order to describe the processes created by computer programs: finalized processes compute functions and must stop by themselves in a finite time. Interaction processes have indefinite duration and install a working space for users until they quit. These interaction processes are often called agents.

But these two views : structural and behavioral are not sufficient to describe completely what are agents and to classify them. So, part 3 presents a genetic or historic view of computing by studying the four forms of causality for functions and for agents. The implementation of the causality make a distinction between mechanical agents and autonomous agents.

From Turing Machines to Interaction Machines

Alan Turing proposed a formal model for the study of functions [Turing, 1936], called Turing Machines (TM). They compute on recursively enumerable sets. They are composed of a tape and a head, which read and write on the tape. The head executes state transitions instructions on the input string.

For a long time, the Church-Turing thesis claimed than TM was the logical model for computers. It was true in the early states of computers, but the material and software evolution needs more powerful models because these machines cannot accept external input while they compute. Peter Wegner proposed [Wegner, 1998] a computational model of Interaction Machines (IM) to express more powerful behavior than calculi for new uses of computers. This evolution in the logical models of computers fit the evolutions of computer software:

Before 1970 two major changes invalidate TM as models of computer software: operating systems are non terminal processes which accept external input while they compute. Files and data bases are permanent tapes which are read and write during computation and which are not reinitialized. Persistent Turing Machines, with two tapes, one for input and one for persistent uses, are an extension of TM to formalize this evolution.

Before1980, a major change invalidate TM and PTM as models: interactive systems accept input streams during computation and produce output streams. The behavior of the system depends of the machine and of the user. Interaction Machines formalize this evolution.

Near 1990, computers are relied on networks, and parallel and distributed systems appear to their users to have undeterministic behavior because

of the multiplicity of hidden interactions. Interaction Machines with multiple interfaces (IMMI) formalize this evolution.
In order to present the results of Peter Wegner, let us explain what are IM and give an example of “pure” IM. The observation of these machines, in an experimental approach, is a mean of validation. Them, in the next part, I will explain why IM can’t be study only in this way.

IM Definition

A process is interactive if it does not control all its inputs and outputs. IM extends TM for interactive processes: TM’inputs are finites and are given before computation (closed systems) but IM’inputs are infinites streams generated by the environment (open systems).
For sequential input stream, with linear pattern of interaction (files, piles, data bases), it is easy to save the idea of Turing Machines and extend them with a persistent tape. The behavior of the machine is sliced and interactions are localized between the slices. So, the theoretical results of TM may be used in each slice. But it is not possible to extend TM in order to preserve theoretical results for multiple interfaces systems, agents with cooperative behavior, reactivity to events. So interactive machines with multiple interfaces where proposed to study the modern behavior of computers.

The smallest IM is echoing

It seams to be a joke that echoing is so important, but it is the smallest pure interactive machine. It can say anything someone says. It can win half the games of chess if it plays two parts in parallel and repeat the play of one great master on the chessboard of the other one.
With a persistent tape, it can learn by observation of human behavior and reuse what happends previously in analog situations. Problem solving by reusability is a new paradigm for IM, when the problem is to have a good interactive behavior in response to a partner.
- Chess || player
- Doctor || patient (Eliza)
- Driver || city

Theoretical Results of P. Wegner

As the input streams of IM are not recursively enumerable, it is easy to

prove the incompleteness of interaction models. So, computing goes beyond logic in providing systematic techniques to solve problems of social life. Interaction machines characterize empirical computer science, and focalize the validation not on theoretical results but on the observation of these machines, as every experimental research do, in physics, biology or psychology [Wegner, 1995]. There were few theoretical work on experimentation in computer science [Nicolle, 2002]. The most important paper before Peter Wegner is from Herbert Simon [Simon, 1995]. An example of the power of this paradigm can be found in the papers of Jacques Tisseau [Tisseau, 2001].
Observation equivalence provides a uniform metric for specifying behavior: two systems S and S' are O-equivalents if they are indistinguishable for a set of observations O. A finer observation reduce abstraction and increase expressiveness. The observation in empirical sciences is the way to gain knowledge.
- Functions are at the first level of the hierarchy of machines : they accept internal observers.
- PTM introduce history and time in calculi : they need external off-line observers.
- IMMI introduce cooperation in calculi : they need external on-line observers, because nothing appends if there is not interaction.

The results of Peter Wegner concern static study of the behavior of modern computer software. The next part of this paper presents a dynamic study of the behavior.

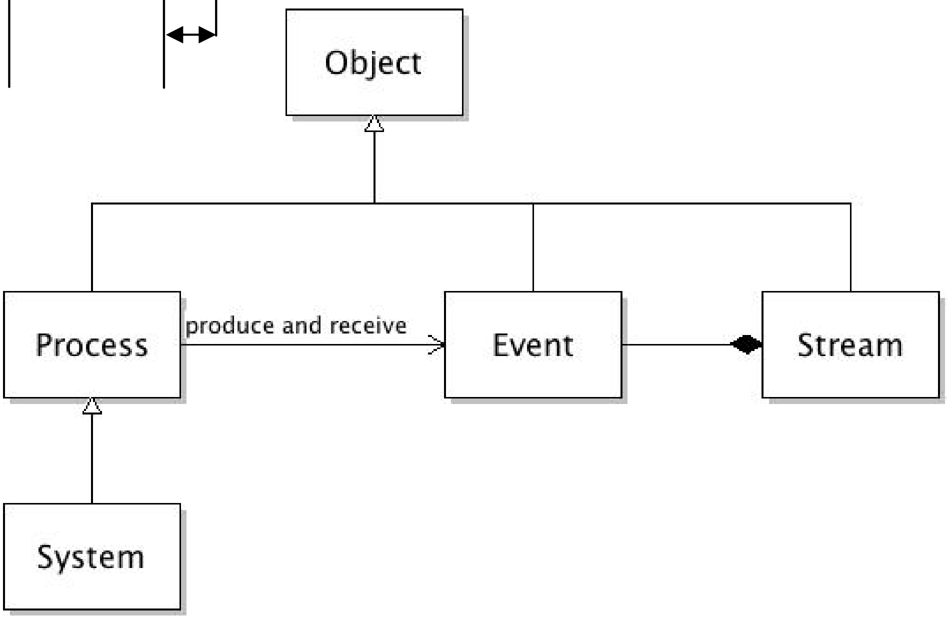
Interaction processes of indefinite duration

Almost interactive computer processes are of potential infinite duration: operating systems, user interfaces, Web services, autonomous agents [Nicolle, 2004]. Interactive indefinite processes can be in interaction with physical process by captors and effectors in order to control technical processes like electric or phone networks, factories, etc. They are in interaction with their human users in a loop of perception and action. They are dynamic and situated in the present time. They control physical actions (open a door) or symbolic actions (ask a question). They modify socio-technical human world by the instrumentation of their communications. We call them agents. The interaction machines

describe these processes in a static way. UML[71] proposes another way to describe modern computer software witch is used in the design and in the documentation of agents. Let us present the main concepts of complex systems by these schemata.

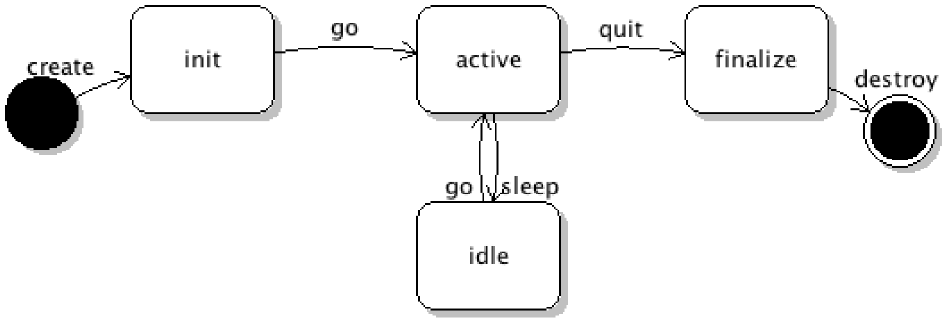
Agents in UML

UML proposes two main kind of schemata for conceptual modeling : structural and relational schemas, and behavioral schemata. The main classes in models of complex computer systems are object, process, system, event and stream. Let us note that System is a Process and is composed of tree or more processes. An event is sent to a process by a user or by another process. For example, in word processing, the user send a stream of events by the keyboard.
The relations between these classes are presented in a UML schemata:



The abstract behavior of an agent is presented in a UML state-transitions schemata:

71 www.uml.org



Transitions between states are crossed when an appropriate event is sent to the agent. All agents have the same state-transitions schemata at the first level. Then, the active state is describe with more details at the second level and is different for each kind of agents.

The activity of agents

First, we distinguished functions and agents in computer software, let us distinguish now two kind of agents, mecanical agents and autonomous agents. Mechanical processes wait for events and treat them (for example, word processing). They repeat a loop of control for technical processes or for the purpose of simulation. They manage data streams (alerts, selective memorization). They never make choices. Autonomous agents have goals and try to achieve them. They perceive their environment and act on it with retrospective effects. They calculate what actions are possible in order to achieve their goal and choice the best one. The more advanced agents evaluate their behavior and learn from their experience. This distinction between mechanical agents and autonomous agents can be explained in terms of causality.

Causality of interactive process

Let us introduce now another way to study agents, where they come from and where they go: the historic or genetic view, whose basis are the causes. Aristotle defined four causes for phenomenons:
- Substantial Cause: phenomenons use substances to appear.
- Formal Cause: shapes constraint moves and the gestalt theory is a modern way to focalize on this question.

- Efficient Cause: physical forces, living forces.
Final Cause: the external reasons of phenomenons.
What does it mean for software agents ?

Substantial causality and digitization

Some phenomenons may exist in many substances. In computational models the independence between phenomenons and their substances is necessary for digitization. Chemical properties like odors and tastes can be simulates in computer systems but not analogically because they are related to substances. They can't be reconstructed from their digitized state, like music, texts or pictures, in a user interface. Only forms or shapes can be digitized in an intrinsic way because they can exist in many substances.

Formal causality and digitization

Formulas, programs, data structures are the formal causes of any computational behavior. Shapes are intrinsically digitizable, so adequate artifacts can exploit the writings of formal causes in order to restore similar phenomenons after their digitization (music, picture). Visual and auditive perception are only perception of shapes and are not destructive like perception of odor and taste.

Efficient causality and modelization

Efficient causality explains the internal reasons of the processes. Physical forces, living forces explain the behavior of natural phenomenons. Psychological or social forces explain the behavior of human phenomenons. There is no equivalence of efficient causality in mathematical modeling. The efficient causality of mathematics is in the human lecture and in their use.
Computers use physical forces in the processor for symbolic uses : they provide an analog of efficient causality. The processor read programs and data and transforms them by physical behavior apply to symbolic use. Instead of mathematical formulas, software models include effective formulas by the mean of this component of causality. The computers produce formal dynamics from writings.

Final causes and modelization

Every artificial process has a final cause which is determined by the goals of the designer. But final cause of technical objects may be transformed by their users. During the process of design of physical artifacts, final causes are transformed in formal causes. In the process of design of software artifacts final causes are often transformed in formal causes and sometimes they are not. We call mechanical processes the processes where final causes are transformed in formal causes. They cannot learn in order to improve their behavior because they forget their purpose.
Explicit final causes characterize autonomous agents: they have goals (viability, problem solving, driving in uncertain environment) and try to realize them. They have to make choices between incompatible goals locally and at the present time. Learning of autonomous agents is managed by final causes and by retrospective effects of actions.

Conclusion

What is new in computer science? When mathematics models by formulas, computer science model by formal dynamics: programs are executed in a physical machine. We can observe them et cooperate with them. Interaction machines and processes of indeterminate duration are the foundation of an empirical computer science. Causal models are more expressive than descriptive models. This science proposes dynamic models of natural or artificial phenomenons. These models are the basis of software agents which create analogs of phenomenons. The implementation of an analog of a phenomenon is a mean to validate the model by the observation of its behavior.

References

A. Nicolle, *Sciences de l'artificiel, modélisation et rationalité*, Revue d'intelligence artificielle, vol. 16, n° 1-2, pp. 63-86, 2002

A. Nicolle, Prolégomènes à une théorie des processus interactifs de durée indéfinie, séminaire Sujet, Théorie et Praxis, MSH Paris, 2004

H. A. Simon, Artificial intelligence : an empirical science AI vol.77, n°1, 1995 p. 95-127, Elsevier

J. Tisseau, *Réalité Virtuelle — autonomie in virtuo*, 2001 www.cerv.fr/~tisseau/doc/hdr/hdrJT.pdf

A. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem, Proc. London Math. Soc., 42:2, 1936, pp. 230-265; A correction, ibid, 43, 1937, pp. 544-546

P. Wegner, *Interaction as a Basis for Empirical Computer Science*, Computing Surveys, March 1995
P. Wegner. Interactive Foundations of Computing. Theoretical Computer Science 192, Feb. 1998

HUMAN MOVEMENT AS A FRAMEWORK FOR UNDERSTANDING INTERACTIONS

Kristine Deray, Faculty of Design, Architecture and Building, University of Technology Sydney, +61 2 95148723, kristine.deray@uts.edu.au

Simeon J. Simoff, Faculty of Information Technology, University of Technology Sydney, +61 2 95141103, simeon@it.uts.edu.au

Introduction

Interactions play central role in our life. We interact with our colleagues and professional team members in the office, with our doctors during the visits to healthcare practices, with our lecturers and classmates in university. Through the interactions with the various computing systems that constitute our contemporary environment we learn about it and achieve our goals. Interactions lie at the very heart of the activities performed in many computerised domains, yet they remain poorly understood. The tendency has been to investigate interactions in terms of the results they produce rather than to show the mechanisms that explain *"how" interactions unfold in time*.

Interactions produce reciprocal effects between parties as perceivable effects. To explain how such effects are constructed, the mechanisms that shape the form and function of the interaction require interpretation and representation. Part of the problem can be related to the insufficient intersection of the language used by the parties, for example, patients and doctors usually bring different background to into the healthcare

interactions they engage. Hence part of the motivation for this research is the development of suitable constructs for facilitating interactions between parties. We look for the ways of identifying such constructs that have been central to the area of embodied cognition (Anderson, 2003), i.e. constructs that are grounded out in terms of human's embodied experience and physical characteristics. Consequently to explain interactions so all parties can derive meaning we need to derive constructs where the vehicle (a medium for communicating expressing or accomplishing something) and the representation, the interaction language, are not separate but are instead constitutive of each other. Following this path we propose that:

Interactions benefit from representations that explain how they unfold in time by referencing the intrinsic dimensions and qualities of the interaction context.

The morphology (form) of interaction is directly related to the function of the interaction context and therefore the coupling form and function generates semantics.

The new representation should be able to facilitate and sustained the "common ground" between parties. Here common ground (a term borrowed from information theory) refers to the knowledge shared by two communicating parties (Berg, 1997, cited by Coiera 2001, p282).

Interactions can be understood as dialogue systems. Dialogue can be perceived as being constituted as content, structure and presence. This indicates suitable representations for interpreting interactions need to support multilevel analysis.

This chapter looks at applying principles of human movement for modeling and understanding interactions in various contexts. In terms of metaphor analysis (Lakoff and Johnson, 1980, 1999), human movement is positioned as the source domain from which suitable constructs for explaining interactions (the target domain) can be derived. We argue that human movement (HM) provides suitable indices for interactions. Movement is the process of change from one position to another. It is a change that provides indices of variability and commonality across contexts that can be utilized to explain what happens in-between positions. If position can be aligned with the notion of a party (a term we take to apply to any entity human or non- human that is involved in an interaction) we can interpret interaction as reciprocal structural and expressive effects between two positions where position references the parties involved in the interaction. Further the systematic basis of HM (as the whole body is the area of concern) naturally provides meaningful relations between part and whole that acknowledge a common reference system. Part- whole relations provide understandings accessible in

different contexts that can be placed to consistently represent interactions at different levels of granularity. Our approach relies on the following assumptions: (i) humans can recognize intuitively language constructs that are based on human movement; and (ii) if the constructs embody meaning then a visual language of forms and functions, derived from these constructs, can provide efficient means for representing interactions consistently at different levels of granularity.

Figure 2 illustrates the transition from the sequence of actions performed by two parties to interactions (the middle space) and then their expression through the movement constructs and corresponding visual representation.

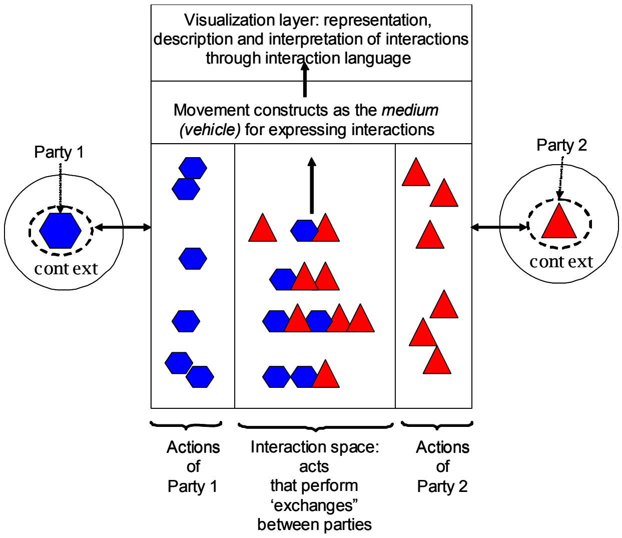


Figure 2. Movement constructs provide a cognitive framework for understandings that facilitate common ground between parties in interactions.

Background

Our motivation to use the domain of human movement as a source for the concepts of our system of reference follows the recognition that humans can intuitively recognize constructs based on human movement, evidenced from the research on kinesthetic thinking/understanding and its ability to support higher-level cognitive processes (Dreyfus, 1992;

Seitz, 1994). The evidence appears from several directions of study. Nicoladis et al. (1999) have studied how children learn to use their gestural and postural abilities to express concepts and ideas through movement. Several researchers suggested that skilled movement is a form of thinking (e.g., Bartlett, 1958, Seitz, 1994, 1996; Fischer and Bidell, 1998). Movement has been found to be predominant in all forms of human intellectual activity (Laban and Lawrence, 1974; Seitz, 2000a, 2000b). Researchers in child psychology (Seitz, 1996; Nicoladis et al., 1999) pointed out that children learn to communicate with gestures before they learn to speak.

Consequently, some cognitive researchers have argued that language understanding and conceptualization are not a result of pure symbolic manipulation, but is a process grounded in bodily interaction with the environment. Glenberg and Kaschak (2003) and Zwaan (2004) provide comprehensive reviews of the works that explain language phenomena in line with the idea of cognition as body-based simulation. Some work in information spaces looked at bodily-kinesthetic skills as a basis for the construction of meaning. For instance, Deray (2000) looked at association between non-verbal movement behaviour and thought processes in the context of human-computer interaction in information search. In a similar simulation context, it is worth noting Bergen et al. (2004), who's approach to the design of a situated grammar references embodied knowledge of perceptual and motor systems that play an important role in higher cognitive functions. Specifically this is in relation to mental imagery, association and memory. Some researchers now support the perspective that many if not "all higher-level cognitive processes are body-based in the sense they make use of (partial) simulations or emulations of sensorimotor processes through the re-activation of neural circuitry that is also active in bodily perception and action" (Svenssson and Ziemke, 2004). The argument is that such constructs embodied in our sensorimotor processes will still reference the physical system they are derived from even when they are linked to abstract concepts. Similarly our understandings of basic spatial concepts are intrinsically linked to how we orientate and move in the physical world. Such reasoning references the experience of the structure of our bodily movement in space.

Several researchers (Barsalou, 1999): (Glenberg and Kaschak, 2002), (Barsalou, 1999): (Barsalou et al 2003) note that it has been proposed that understanding a piece of language entails internal simulation and/or mental imagery that provides access to the same neural structures "that would be involved in perceiving the precepts or performing the mental

actions described”. This is in line with recent research in neuroscience that indicates a neural basis for embodied understandings.

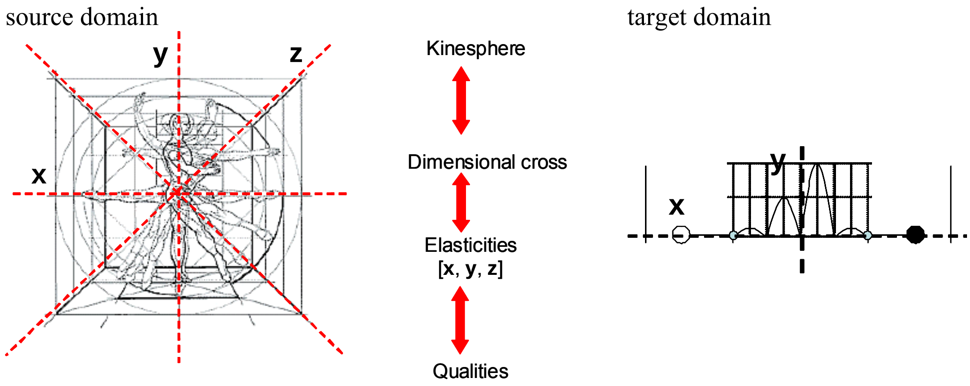
The argument that kinaesthetic thinking, kinaesthetic logic, supported also by the linguistic view of the mechanisms behind metaphors (Lakoff and Johnson 1980, 1999;), is fundamental to human thinking, then, has value, inferring that the concepts of human movement could provide consistent and sustainable basis for developing means for representing, expressing and analysing interactions. Further, in Section 3, we present the proposed methodology for designing visual languages that utilise aspects of human movement.

Methodology

In this section we consider contact improvisation and movement observation science to derive systems of reference that can frame an understanding of interactions. To derive suitable formalisms we adapt the approach of movement observation science and apply it to contact improvisation - a style of post modern dance, whose techniques unfold in a manner similar to conversation. In contact improvisation there are no defined movement types, which define a “formal grammar”, for instance, like in classical ballet. Interaction between parties evolves as movement actions structurally determined that generate some reciprocal effects between parties (performers). Contact improvisation provides a rich source domain for understanding how interactions unfold by providing constructs that model intrinsic parameters of human movement as relations between parties in the “dialogue”. Practitioners in the area of movement observation science have derived movement notational systems, designed to record human movement in symbolic form. One of these systems – Labanotoation, developed in the 1920’s by Rudolf Laban’s team, and the subsequently developed Laban Movement Analysis with its Effort and Shape components, provide us with valuable formalisms for extracting our movement constructs (Newlove, 2001; Newlove and Dalby, 2004). These models relate two components or frames of reference of human movement: (i) body position - the place of the body in space, and; (ii) body dynamics - the motion that causes and expresses change from one position of the body to another. Further in the chapter we focus on the elicitation of the constructs that are based on human movement and their utilization for representing interactions.

Systems of reference for interactions

We drive systems of reference from the domain of human movement by understandings based on the whole human body and the structural relations that mark movement from one position to another. From this interpretation constructs derived from a systematic approach can be related and indexed. We take three levels - surface, middle and deep, for the analysis, as illustrated in Figure 3. The constructs describe both unchanging and changing features of the systems of reference at different levels. Four levels are discussed as follows: (i) kinesphere; (ii) dimensional cross; (iii) elasticities; and (iv) qualities. Figure 3a represents the *kinesphere* of the body - a spatial framework that describes the total volume of possible moves of the components of the human body, illustrating its dynamic boundaries.



a. systems of reference of human movement (the source domain) (adapted from (Kaiser, 1999)

b. the different levels of the systems of reference

c. levels in the interaction domain modeled from constructs and relations derived from source domain

Figure 3. From systems of reference to representation of interactions

The *dimensional cross* is the term given for the spatial dimensions in which we move: upwards and downwards, backwards and forward and from side to side (this is Laban's term as noted by Newlove and Dalby, 2004, p49). These spatial dimensions intersect at the centre of the kinesphere that corresponds to the centre of the human body. The dimensional cross can be depicted graphically by corresponding axes, that is, x axis represents side to side, y axis represents up/down; and z axis corresponds to forward/backward. Each dimension can be perceived as a continuum of points with two polarities one at each end of the axis. This configuration provides a reference system by which the orientation of the kinesphere can be indexed. These three dimensions of the cross link to the *elasticities*, which are structured by the three anatomical actions of the body commonly referred to as contract (or flex), extend and rotate. The anatomical actions work along side of the three dimensions in the dimensional cross. We refer to these movement aspects as elasticities: the contraction and extension elasticity (corresponding to side to side movement along x axis; the rising and sinking elasticity corresponding to up and down movement along y axis; and the rotate and tilt elasticity corresponding to rotational forward and backward movement along the saggital axis z. The elasticities by their behaviour, individually or in combination, express shaping affinities on each dimension of the body referencing the anatomical actions. The relation

between the dimensions of the body and shaping affinities has been noted in the work of (Zhao et al., 2000).

Qualities derived from the human movement conceptualization, based on Laban Movement Analysis and the Effort Shape framework, provides us with the conceptual basis for the mapping from the source to the target domain in our framework. In this chapter we consider four basic elements Laban associated with effort shape, namely (tension) flow, weight, time and space. Laban Movement Analysis links these concepts through the notion of effort, which describes how the body concentrates its exertion while performing movements. The resultant shape of movement is a direct outcome of the parameters of effort that give form to the movement. Effort has been compared to dynamic terms in other domains such as music – and of note for our approach has been considered to explain how a piece of music is performed. Effort interpreted in this manner describes the unfolding of action in a particular context. The parameters derived from these qualities are summarized in Table 1. Each motion quality is a continuum between two polarities that describe the extreme values for that continuum. We argue that effort and the resultant shape of movement provide distinct parameters that can be mapped to communicate relations derived from the target domain – the interaction context of interest.

Table 1. Qualities in the source domain derived from the motion factors and Effort elements of the Laban Movement Analysis.

Motion qualities	**Effort elements**	**The role of the concept in human movement**
Flow	Free Bound	Describes whether a movement is bound or relatively free with respect to the human body
Weight	Light Strong	Describes the “easiness” of movement, the quality of lightsomeness or forcefulness of movement
Time	Sustained Sudden	The length of a movement, or movement phrase: describes how the movement was communicated in, for instance, in a sudden or sustained manner
Space	Flexible Rigid	The spatial focus of movement: describes spatial relatedness of movement elements to a single focus point or being divided amongst several foci

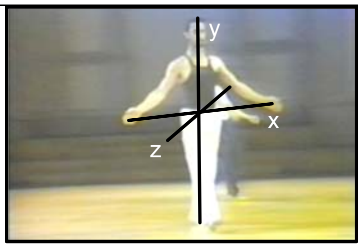
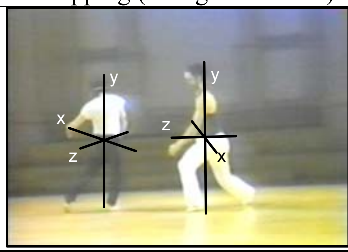
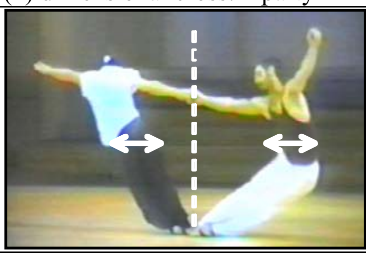
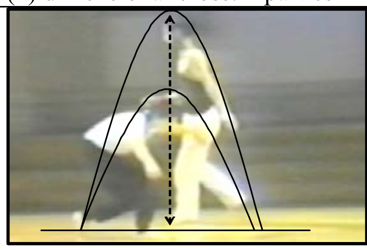
Each motion feature affects the shape of movement that in turn is correlated to the amount of effort the movement requires. The effort elements, what we have named qualities, can be combined in infinite number of ways to express a trajectory of human movement. The levels provide systems of reference that can explain how movement happens. The four levels together provide means for systems of reference that can be indexed top down or bottom up to give multilevel analysis of both position and dynamics. Follows application of the described systems of reference and levels to instances in the domain of dance- movement, that is, contact improvisation.

Lastly Figure 3c represents the “kinesphere” of the target space, interactions. As in the source domain levels of analysis can be constructed that explain the behaviour and relations between parts. In Figure 3c the kinesphere is significantly different form in Figure 3a. At this stage of the modeling the “dimensional cross” in the target domain has only two axes as the forward backward axis is currently not active.
Further, we formalize these concepts and utilize them in our visual language, presented in Section 6.

Applying movement observation science to derive systems of reference

Figure 4 demonstrates the derivation of the key movement constructs from examples of contact improvisation, in particular, describing instances of: (a) the kinesphere; (b) the dimensional cross; (c) the three elasticities with the corresponding shaping affinities; and (d) the qualities of effort shape. The series of frames describes the “dialogue” between two parties. By applying the systematic analysis based on the systems of references, described in Section 4, the interaction can be interpreted through the key constructs discussed. Each interaction has a kinesphere of possible relations that are constructed as the superposition of the kinespheres of the individual parties engaged. Frame (a-1) illustrates individual kinesphere, defined by the limits of the limbs. Frame (a-2) shows variation provided by the overlap of the kinepheres of the two parties. The dimensional cross assists in understanding the orientation of the body in space. Frame (b-1) illustrates the dimensional cross of an individual. Frame (b-2) illustrates how the orientation of the dimensional cross indicates the relations between the individual kinespheres. In (c) four instances of the elasticities are given. In frame (c-1) the elasticity of contraction-extension is clearly stated as the two parties pull away from each other to maximum length. The affinity to the horizontal dimension is clear. In frame (c-2) the elasticity of rising-sinking can be understood by comparing the two body poses. The figure on the left is sinking as she curves down to the floor. The figure on the right retains a vertical position and can be considered to express rising. In frames (c-3) and (c-4) the elasticity of rotation-tilt is expressed with (c-3) showing the counter placement of components of the body rotating forward and backward, while (c-4) shows the rotation of one party in the air by making use of the torque of the standing rotating party. In (d) instances of the four qualities of effort are given. In frame (d-1) space quality describes the relatedness

of the two parties. The figure on the left has a “single focus” (in terms of Laban this means that the body elements are aligned in one direction). The figure on the right demonstrates more than one focus which is perceived as a more flexible state of the body. In frame (d-2) the weight quality is given the value of lightness indexed to the figure on the right as she jumps into the air while the figure on the right is neither forceful or light but sits in a middle neutral range. In frame (d-3) flow is shown with a value of bound which is given by the crossing of the legs that limits movement in this instance. Note however the arms are free being open and un-crossed. This often occurs in movement as the anatomy of the human body supports independence of parts in space. Lastly in frame (d-4) sudden time is expressed by the jump of the party into the kinesphere of the receiving party. The dotted arrow indicates the trajectory that body took in this instance.

Construct	Frame example	Frame example
(a) Kinesphere		
	(1) kinesphere of one party	(2) kinesphere of two parties overlapping (changes relations)
(b) Dimensional cross		
	(1) dimensional cross: 1 party	(2) dimensional cross: 2 parties
(c) Elasticities		
	(1) contraction-extension shaping affinity to the horizontal x axis	(2) rising sinking elasticity: shaping affinity to the vertical y

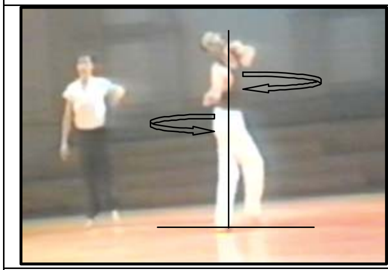
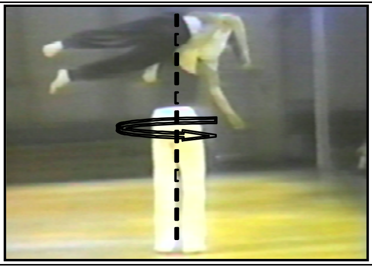
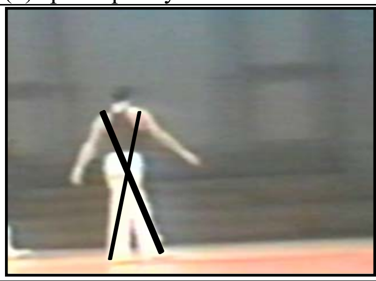
		axis
		
	(3) rotation/twisting elasticity: shaping affinity to retreating-advancing.	(4) rotation of one party provided by the other
(d) Qualities		
	(1) space quality	(2) weight quality
		
	(3) flow quality	(4) time quality

Figure 4. Movement constructs derived from examples of contact improvisation using movement observation science approach (the source of contact improvisation is the Videocoda performance by Steve Paxton /Nancy Stark- Smith 1983).

The above is a very simple overview of the constructs discussed and is in no way exhaustive. Although not indicated in these instances all movement has a proceeding and proceeding phrase that gives a sequence of actions and a chain of relations. The last instance, the jump into the space of the other dancer, is linked structurally and expressively to the momentum the dancer has derived from the immediate preceding movements. Figure 4 below describes an example of a sequence of movement where the relations between parties are linked into a chain of kinematic actions until the kinespheres of the parties separate (i) and the

‘interaction set” is finished. Similar to Figure 3, we use graphical indicators to describe, (a) the kinesphere, the elasticity contract-extend (e), (f), (g) and the elasticity rotate-tilt in (e), (f) and (g). In the next section we briefly describe how we derive the language primitives. The complete description of the language with detailed examples of its constructs and applications is presented in Deray and Simoff, (2006).

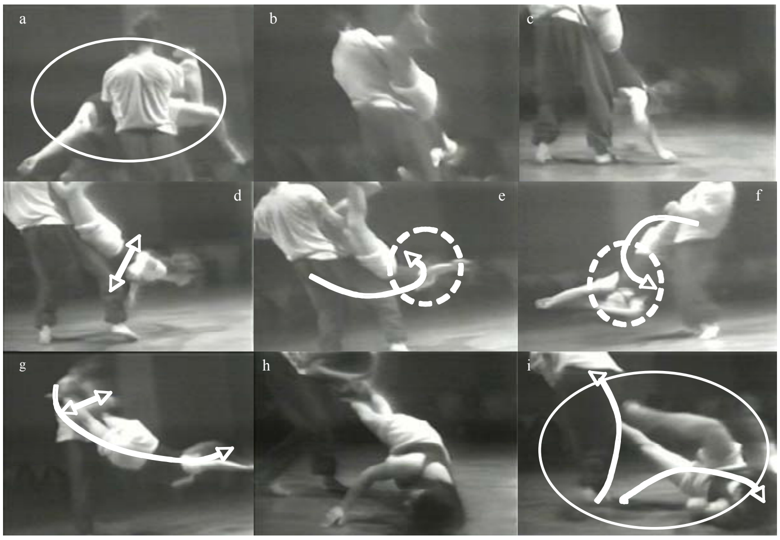


Figure 5. Unfolding of interaction in CI between two parties (the source of contact improvisation materials is Videocoda 1973)

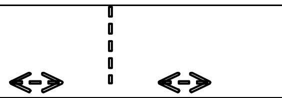
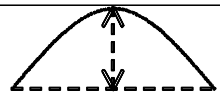
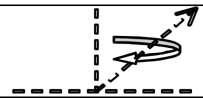
Deriving visual language primitives and expressions

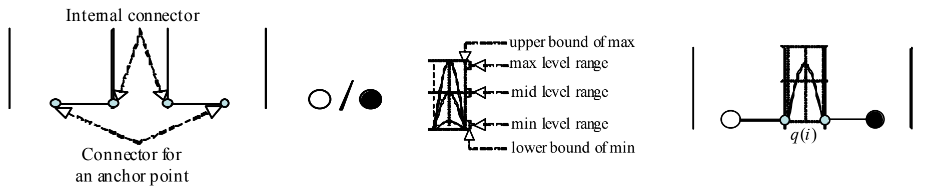
We apply the guidelines on the forms and behaviour of the visual elements, derived from the movement constructs discussed in sections 4 and 5, to the target domain in order to derive a language for expressing, representing and explaining interactions.

Table 2 presents the slice of the mapping between movement constructs and interaction features along the elasticities illustrated in Figure 4. The shaping affinities of the elasticities are linked, through their semantics, to the interaction features. The mapping provides the basis for constructing the visual primitives of a language for expression of interactions. The semantics of the movement constructs also defines the constraints and affordances allowed.

Figure 6 shows the basic visual primitives of a language for representing interactions, whose shape and behaviour has been derived from the movement constructs.

Table 2. Constructing the forms of the basic elasticities (based on the corresponding anatomical actions)

Elasticity	Contraction and Extension	Rising and Sinking	Rotate and Tilt
Diagrammatic representation			
Semantics	The amplitude on the horizontal axis indicates the strength (weakness) of interaction. It models strength as states of attraction, repulsion or stable behaviour.	The amplitude on the vertical axis indicates the elastcities that stretch up and sink down.	The angle of rotation on the sagittal axes. Measured by angle increment.
Interaction features	Describes the *intensity* of the interaction.	Describes *flow* and *effort* of the interaction.	Describes *associations* between interactions over time



(a) body element (b) anchor point (c) effort–shape element (d) interaction element

Figure 6. Primitives for representing interactions.

The current version of the language considers two of the elasticities presented in

Table 2 – the elasticities of “Contraction and Extension” and “Rising and Sinking”. Each of these elasticities may operate with one or more qualities. Currently, the “Contraction and Extension” elasticity operates with a single quality, and the “Rising and Sinking” elasticity – with four nominated qualities.

The elasticities provide the semantics that specifies the behaviour of the visual primitives so that their form changes according to the movement constructs. The body element in Figure 6(a) indicates the strength of the interaction. This elasticity operates in horizontal dimension and as the interactions that we consider are between pair of participants, the body element unit consists of two “track segments” (fields) of equal width, where each quality of the “Contraction and Extension” elasticity is represented as two horizontal “arms” (line segments) - one for each participant. An anchor point, shown in Figure 6(b), represents a party involved in the interaction with static (passive) (○) or dynamic (pro-active) (●) behaviour. The internal connectors link the body element to the elements of the “Rising and Sinking” elasticity. We label as an effort-shape element the element that visually represents the qualities of the “Rising and Sinking” elasticity Figure 6(c). The effort shape element unit describes the qualities of interaction. The four qualities derived from the vehicle, human movement (see Table 1), are flow (values of bound/ free), exertion (values of light/strong), transition, (values of sudden/ sustained) control (values of rigid/flexible). An effort shape element comprises from one or more effort shape units with two connection points at the end. The range of each effort shape unit can map a continuous or discrete value. A collection that includes a flow element, two body elements and two anchor points constitutes an interaction modeling element, shown Figure 6(d). The interaction modeling element can be viewed as an “interaction kinesphere” as it behavioural principles follow the principles of the kinesphere in human movement. At present, the elasticities operate in a two - dimensional reference system.

Conclusions and future work

In this chapter we demonstrated that human movement provides concepts that can be applied for modeling and understanding interactions. The theoretical framework underpinning these concepts is derived from movement observation science, specifically, the Labanotation and Laban Movement Analysis (Guest, 2005). We

presented the methodology for deriving means for expressing interactions through the application of human movement constructs. We argue that interactions as sets of interconnected actions that unfold in time between involved parties can be interpreted by representations, which by their nature are intrinsic to the domain of human movement. Such constructs express both relations of position and dynamics between parties. The approach utilizes human observation science and contact improvisation to derive the constructs that are suitable for description and explanation of interaction. Currently, the visual language for interaction representation utilizes only two of the three elasticities. The future work is focused on the utilization of the third elasticity and the granularity of each elasticity, i.e. the number of qualities through which each elasticity operates.

References

Anderson, M. L. (2003). Embodied cognition: a field guide. *Artificial Intelligence*, 149, 91-130

Bartlett, F. C. (1958). *Thinking: An experimental and social study*, New York, Basic Books.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioural and Brian Sciences*, **22**, pp. 577-609.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., and Ruppert, J. M. (2003). Social embodiment. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*, **43**, San Diego, Academic Press.

Bergen, B., N. Chang, et al. (2004). Simulated Action in an embodied construction grammar. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, Chicago, IL.

Coiera, E. (2001). Mediated Agent Interaction, In *Proceedings of the 8th Conference on Artificial Intelligence in Medicine Europe*, AIME 2001.

Deray, K. (2000): "Through "Plane": A kinetic model for information retrieval. *Proceedings of the Eleventh Australasian Conference on Information Systems ACIS 2000.*

Deray, K. and Simoff, S. J. Designing a visual language for interaction representation, based on principles of human movement, in Ferri, F. (ed) *Visual Languages for Interactive Computing: Definitions and Formalizations,* in publication

Dreyfus, H. L. (1992) *What Computers Still Can't Do: A Critique of Artificial Reason*, Cambridge, MIT Press

Fischer, K. W. and Bidell, T. R. (1998). Dynamic development of psychological structures in action and thought. In Bidell, T. R. and Damon, D. W. (eds), *Handbook of*

Child Psychology: *Vol. 1: Theoretical Models of Human Development*. New York, John Wiley and Sons, pp. 467-561.

Glenberg, A. M., Kaschak, M. P. (2003) The body's contribution to language. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*, **43**, San Diego: Academic Press.

Guest, A. H. (2005): *Labanotation: The System of Analyzing and Recording Movement*. Routledge, New York.

Laban, R. and Lawrence, F. C. (1974). *Effort: Economy of Human Movement* (2nd ed.), Boston, Plays Inc.

Lakoff, G., Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G., Johnson, M. (1999). *Philosophy in the Flesh*. New York: Basic Books.

Newlove, J. and Dalby, J., (2004). *Laban for All,* London, Nick Hern

Newlove, J. (2001). *Laban for Actors and Dancers,* London, Nick Hern

Nicoladis, E., Mayberry, R. L., et al. (1999). Gesture and bilingual development. *Developmental Psychology*, **35** (2), pp. 163-174.

Paxton, S. and Stark Smith, N, (1983) *Fall After Newton,* Videocoda/ Contact Colloboration

Seitz, J. A. (1994). Thinking kinesically: Theory and practice. In the *Proceedings of the 24th Annual Symposium of the Jean Piaget Society*, Chicago, IL, Jean Piaget Society.

Seitz, J. A. (1996). Developmentally appropriate practice. *Journal of Education and Family Review* **3** (5), pp. 7-9

Seitz, J. A. (2000a). The embodied self. In *Proceedings of the 30th Annual Symposium of the Jean Piaget Society*, Montreal, Canada. .

Seitz, J. A. (2000b). Embodied cognition. In *Proceedings of the 12th Annual Convention of the American Psychological Society*, Miami, FL

Svenssson, H. Ziemke, T. (2004). Making Sense of Embodiment: Simulation Theories and the Sharing of Circuitry Between Sensorimotor and Cognitive Processes, *CogSci 2004,* Chicago

Zhao et al (2000) Interpreting Movement Manner, *Proceedings of Computer Animation,* 112-120

Zwaan, R. A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In: B.H. Ross (Ed.), *The Psychology of Learning and Motivation*, **44**, New York: Academic Press.

TEACHING PHILOSOPHY USING HYPERTEXT

THE LOG EXPERIMENT

MARYVONNE LONGEART

Professeur agrégée de Philosophie, Lycée de Villard-de-Lans; Professeur-concepteur au LOG (Lycée Ouvert de Grenoble); Membre du groupe de recherche RESEAU.RAISON, Université Laval, Québec

Tel.: 04 76 95 26 53; e-mail: longeart.roth@wanadoo.fr

RÉMI CLOT-GOUDARD

Professeur agrégé de Philosophie, Académie de Grenoble; Professeur-concepteur au LOG (Lycée Ouvert de Grenoble); Membre du groupe de recherche RESEAU.RAISON, Université Laval, Québec

Tel.: 06 77 27 52 45; e-mail: r.clot-goudard@netcourrier.com

Abstract: This article addresses the pedagogical problem of the non-linear reading of philosophical texts induced by hypertext links in a computerized knowledge base used for computer aided teaching of philosophy. This paper should be considered as an experiment report. We will first present a general description of the LOG philosophy text base and then discuss the rationality at work in hypertext links, both from the point of view of the designer of the base and of its users.

Introduction

In teaching philosophy using computer-based tools, an essential feature is the possibility of making what Woods [1] used to call "the kinds of subtle distinctions that can be made by people in conceptualizing complex ideas". The whole enterprise could be viewed as a form of "applied philosophical investigation".

Instead of a purely theoretical discussion of the issue, we present selected results of a practical experiment, which has in itself some theoretical significance. We show how one can teach philosophy using the LOG system, an experimental e-learning tool we developed to help special-need students who cannot attend regular classes. The hypothesis is that such a computer-based textbook in philosophy can give insight into the conceptual structure of philosophical problems. In this paper, we intentionally avoided technical issues related to either the design of the system or History of Philosophy to focus on the benefit to enhance traditional humanistic teaching with tools and methods developed in Computer Science and the need to further develop more appropriate tools.

This experiment must be evaluated in light of the whole problem of computer-aided learning. The central point of this presentation is to show that hypertext provides an adequate vehicle for investigating conceptual interrelationships between philosophical texts. We will not undertake a substantial analysis of specific philosophical theories – a goal beyond the scope of this report.

We will first present a general description of the LOG philosophy text base and then discuss the rationality at work in hypertext links, both from the point of view of the designer of the base and of its users.

The LOG system: A brief Description

LOG, Lycée Ouvert de Grenoble (Grenoble Open High School), is a virtual teaching environment intended to assist high school students unable to attend regular classes: athletes or musicians involved in high-level competition or intensive practice, students suffering from illness or handicap… LOG has no building or classroom: students and teachers using the system belong to regular teaching institutions in different locations throughout the Académie de Grenoble. Teachers in the different

fields of the curriculum, design the teaching material: Literature, Maths, Biology, History, Geography, Greek and Latin, Philosophy. The LOG resources are accessible through the Internet and are designed to be used under the supervision of a teacher who acts as a tutor.

Philosophy has been introduced in LOG only two years ago with very limited resources. Designing the teaching material is a very slow process and the system is still under construction. But although the experiment in philosophy is still at an early stage of development, some interesting observations have been made on the kind of rational thinking at works in the use of the tool.

The content of the LOG base in philosophy is constrained by the official program for Grade Twelve (classes terminales) in the French secondary school system. The program consists of a list of twenty-eight notions that make the core of the program, a list of conceptual distinctions students should learn to master and a list of philosophers. Conceptual distinctions and authors are not studied for there own sake. The philosophy class is an introduction to philosophical analysis centred on the study of philosophical problems related to the notions suggested in the program list.

As shown in *Figure 1*, we decided, as designers of the tool, to divide the study of each notion into four modules:

- *Etonnement* is an introduction to philosophical questioning using one or several non-technical documents.
- *Conversation et débats entre les textes* is the further exploration of related philosophical problems through a network of texts related by hyperlinks. (This network is the core of the tool.)
- *Cheminement* proposes a guided study of a more substantial piece of text — a book, a chapter of a book or an article. The study, at length, of at least one work by a philosopher is compulsory. The works studied during the school year are to be presented by students at the oral test of the Baccalauréat. The choice of a specific work is left to the teacher, but it has to be written by one of the philosophers in the official program list. We selected one work per notion, providing online teaching material for each work (mainly through hyperlinks).
- *Methode* provides various exercises in conceptual analysis as well as reading and writing philosophical texts. This module helps students meet the requirements of the written exam of the French Baccalauréat.

Figure 1: Index Page for the Notion "Culture"

It has been our methodological choice, as designers of the tool, to put the emphasis on texts written by philosophers and organised into a network by means of hypertext links. Links can be classified into a typology according to the function they serve in the base. *Figure 2* shows an example of a text by Montaigne along with the different types of links provided to the reader:

- *Navigation links* are means to manoeuvre inside the base: going from one module, or one notion, to another, choosing an item in a list or a menu.
- *Annex links* lead to a new page providing additional information on a subject: a presentation of the author, the list of other texts by the author available in the base, or a more technical approach to a conceptual distinction.

- *Note links* play the same role as footnotes, providing instructions, references or explanations.
- *Hypertext links* as such lead to another text in or outside the base. They are the arcs between text nodes that build the text network.

Hypertext links give the user extensive flexibility in the links interpretation, at the cost of clear semantics for the resulting net. The interpretation of a link is largely left to the “reader” of the hypertext. This feature could be seen as a strength or a weakness depending on whether one wants to emphasize non-ambiguity or flexibility.

Figure 2 shows an example of text and links in the LOG base. Clicking on a navigational link leads to another part of the base (the notion index page or the base index page, for example); clicking on the book icon leads to a list of other texts by Montaigne in the base; clicking on Montaigne’s portraits leads to information on the author; clicking on a link inside the main text leads to another text, for example, here, a text by Lévi-Strauss on the same issue; placing the pointer over a note link prompt additional information to appear in margin as shown here: Lévi-Strauss’s quote “ Le barbare est celui qui croit en la barbarie” appears in margin when the pointer is over Montaigne’s text “chacun appelle barbare ce qui n’est pas de son usage”.

Figure 2: An Example of Text and Links

Hypertext and rationality

According to John Woods [1], two goals have been associated with the choice of a knowledge representation tool, and this applies to the choice of a teaching tool as well:

- Expressive adequacy: what the tool can indeed represent about the subject matter, i.e. what distinction it can help draw and what it should leave to the student to figure out.

- Notational efficiency: the representational efficiency for various kinds of teaching material along with conciseness of representation and ease of manipulation.

A good tool should address both simultaneously.

What are we entitled to expect from the use of digital resources exploiting the power of hypertext links? Is there a *specific difference* of hyperlinks, properties that belong to them exclusively and give them a *proper* pedagogical virtu in teaching philosophy ?

Very strong objections have been raised against the use of computers in philosophy in general and in teaching philosophy in particular. Most philosophy teachers are, at best, reluctant or dubitative if not grudgingly hostile to it. Dubitative: what can we gain by working with computers, if anything we do with them can be easily accomplished using regular printed material ? Hostile: computers and hypertext, far from developing the mastering of rational thinking, reinforce divestment, conceptual confusion and eclecticism. Of course, students do like playing with computers. The screen and the virtual manipulation of its display fascinate them. Hence, the text disappears behind its own image, becoming nothing more than an image like any other image appearing on the screen, and the philosophical work dissolves into idolatry of the tool. It is a serious objection and there is some truth in it. According to Jacques Ellul [3], treating machines in general, as idols is a recurrent attitude induced by technique. This alone doesn't prove that any use of a computer is incompatible with a philosophical attitude. However, given the conscious and systematic way in which a philosophy is developed and constructed, the process of conceptualization inherent to any philosophical enterprise is not easily grasped by the informal and often unpredictable manoeuvring of hypertext links between atomised parcels of text. In fact, dispersion and confusion are the strongest objections. It means that Hypertext links are obstacles to the learning and mastering of rationality standards. Rationality is here at stake. But what *kind* of rationality ?

It seems that the objection relies on two presuppositions regarding the nature of the pedagogical process and what it implies in term of rationality:

1. The first presupposition, we could call the “Cartesian Predicament” according to which the rational mind, associated with what he calls “Natural Light” or “Reason” is a *planning* mind [2].
2. The second presupposition, we could call the “Platonic Predicament” according to which philosophical training can only be done through person to person relationship

The Cartesian predicament

According to Descartes, there are two aspects to the power of Natural Light: the power of discerning the true from the false and the power of reasoning, that is to say of deriving truths from one another, the model of which is deduction (going from a logically first principle to its consequence). Since childhood, prejudices, opinions and teachers that bend the will away from it dim this Natural Light. This is why a *method* is required, i.e. an education of the will:

> “ Le bon sens est la chose du monde la mieux partagée car chacun pense en être si bien pourvu, que ceux même qui sont les plus difficiles à contenter en toute autre chose, n’ont point coutume d’en désirer plus qu’ils en ont. En quoi il n’est pas vraisemblable que tous se trompent; mais plutôt cela témoigne que la puissance de bien juger, et distinguer le vrai d’avec le faux, qui est proprement ce qu’on appelle le bon sens ou raison, est naturellement égale en tous les homes; et ainsi, que la diversité de nos opinions ne vient pas de ce que les uns sont plus raisonnables que les autres, mais seulement de ce que nous conduisons nos pensées par diverses voies, et ne considérons pas les mêmes choses. Car ce n’est pas assez d’avoir l’esprit bon, mais le principal est de l’appliquer bien.” Descartes, Discours de la méthode, Première partie.

The four rules Descartes found adequate to express his method are abstract prescriptions to which the will must comply in order to enable proper use of the “*bon sens*” or reason the human being is naturally gifted with. Philosophy would be nothing else than an educative process by which the perverted will submits back to reason. Teaching philosophy boils down to the taming of the will. Rationality is nothing else. The philosophical ideal pursued is the mastering of the self and his discourse, guided by recognition of clear and distinct perceptions. What is important for those who intend to learn the philosophical process is to proceed by analysis and deduction, justifying all steps along the way.
In that perspective, using hypertext links may appear a dubious auxiliary falling short of helping the student to take full control of the reasoning

process and deliberately exercise his natural ability to tell apart the true from the false.
However, no matter how serious this objection might be, hypertext and hyperlinks may turn out to be, although in a somewhat non-Cartesian way, a philosophical exercise of reason conceived as something else than a pure logical capacity of deduction.

The Platonic Predicament

The platonic presupposition holds that philosophical training needs to be an interpersonal relationship. Rational thinking requires our ideas to be confronted with another rational agent's ideas. The dynamic of this confrontation, or dialogue, relies on the obligation to convince by agreeing on terms, exchanging arguments and using demonstration. Refutation is a major aspect of the process: to show that a proposition is incompatible with other propositions put forward by the opponent in the dialogue. Coherence is the ultimate requirement in any rational thinking.
The objection is a semantic one: a machine cannot be an adequate protagonist in a dialogue because the onus couldn't be upon it to do the refutation, considering its indifference to the meaning of what is said or done. A machine cannot detect ambiguities or conceptual confusions leading to incoherence. It cannot reinforce the requirement for terms definition and conceptual distinction.
However, this objection is limited. Even if we agree that a person to person relationship is a necessary condition to philosophical training, it doesn't implies that nothing else can help in the process.
The real question is in fact whether or not hypertext can be a way to exercise the student's reason and to what extent this exercise is a philosophical one.

Hypertext and the Intentional Stance

How can we characterise the distinctive feature of hypertext ? What does it require from the students ?

On the one hand, one should take into account that the existence of a link is not a matter of chance. Links in the LOG base, for example, are the effect of the designers' decision to actually materialise a relationship. The resulting network is an organised whole, a cosmos, where everything, in principle, takes place according to a leading intention.

On the other hand, even the most naïve user assumes the artificial status of the base, i.e. the fact that it is the work of a human agent. The artificiality of the base, so to speak, and its explicitly pedagogical destination guarantee the rationality of its construction and, ipso facto, of its links. Links in the base are there for a purpose. Vis-à-vis a link, the user is in a position to interpret comparable to the one we are in when we face what we know or believe to be a rational agent capable of purpose oriented behaviour. It is what Daniel Dennett refers to as “the intentional stance” [4].

The user of hypertext spontaneously believes that a link has a *raison d’être*. However, this justification may not be explicitly stated. It *could* be (the cursor, positioned over a link, sometimes displays a message explaining the meaning of the link), but if it is not the case, it is left to the user to figure out by himself what could be the *raison d’être* of the link. In order to do so, he must activate the link and think about what is intended by the designer who made the link available. In doing so, the user applies Quine’s “principle of charity”.

A hyperlink is the materialisation of a conceptual link established by the designer of the base. It reflects the designer’s own understanding of conceptual problems to which the base is supposed to introduce its reader. The meaning of a link might not be obvious, but the reader is entitled to postulate that there is a meaning. It is his responsibility to try to reconstruct the suggested conceptual link through the activation of the material hyperlink.

In other words, the *specific difference* of hyperlink is to materialise for its user an *opportunity* to recreate a *conceptual space*. This is what intelligence is all about: *intelligere* is to tie links between previously unconnected items. Of course, there is no guaranty that the resulting conceptual space will coincide with the intended conceptual space of the designer. The meaning of a link can be blurred or misinterpreted.

The user’s postulate, that hypertext links do have an immanent rationality to be reconstructed, mirror the designer’s postulate of a rational user able to understand the *instrumentality* of a link, its insertion within a means/end relation to the benefit of reflective thinking.

Hence, the rationality involved in hypertext is pragmatic.

Conclusion

Coming back to the two objections above, we are now in a position to conclude.
Considering the “Platonic Predicament”, we have seen that even if the interaction between the user and the LOG base implemented on a computer is a “dialogue” only in a metaphorical sense, this interaction is not deprived of any pedagogical virtu: it induces questioning and enhances the capacity to “make sense”.
Considering the “Cartesian Predicament”, we have shown that the rationality of the users is required. But it is, so to speak, his “practical” reason: his ability to grasp mean/end relationship, subordination orders in the concrete interaction with the base as *artefact*.

From a practical point of view, the tool discussed turns out to be useful for philosophical investigation for several reasons. At the very least, it can make a student’s life easier by pointing out, through hypertext links, some relationships in a text and between texts that might not be so readily discernible with traditional reading methods. As such, it can serve as a good pedagogical tool at the introductory level. Furthermore, students might benefit from the interactive display of conceptual relationships. A deeper understanding of the structure of a conceptual system might be gained and this could lead readers of philosophical texts to understand them in new ways. But we did point out certain limitations of hypertext. An obvious challenge must be overcome for this approach to be successful. In order to illuminate philosophical texts by computer-based hypertext network linkage, it is required to cast those links into a form not only the system but also the students can handle properly. However, in doing so, we clearly run the risk of introducing some distortion or misunderstanding. One should not forget, however, that the LOG base is not intended for use by students in complete autonomy. The teacher remains the ultimate *garde-fou*.

References

Woods, W. A. “What’s Important About Knowledge Representation?” IEEE Computer, 16, 10, 1987, 22-27.
Suchman, L., *Plans and Situated Actions*, Cambridge, CUP, 1987.
Ellul, J., *Le Bluff technologique*, Paris, Hachette, 1988.
Dennett, D., *The intentional Stance*, Cambridge, MIT Press, 1983.

KNOWLEDGE AS COMPUTATION IN VIVO

SEMANTICS VS. PRAGMATICS AS TRUTH VS. MEANING

GORDANA DODIG-CRNKOVIC

Department of Computer Science and Electronics

Mälardalen University

Västerås, Sweden. *gordana.dodig-crnkovic@mdh.se*

Abstract. Following the worldwide increase in communications through computer networking, not only economies, entertainment, and arts but also research and education are transforming into global systems. Attempts to automate knowledge discovery and enable the communication between computerized knowledge bases encounter the problem of the incompatibility of syntactically identical expressions of different semantic and pragmatic provenance. Coming from different universes, terms with the same spelling may have a continuum of meanings. The formalization problem is related to the characteristics of the natural language semantic continuum. The human brain has through its evolution developed the capability to communicate via natural languages. We need computers able to communicate in similar, more flexible ways, which calls for a new and broader understanding far beyond the limits of formal axiomatic reasoning that characterize the Turing machine paradigm. This paper arguments for the need of a new approach to the ideas of truth and meaning based on logical pluralism, as a consequence of the new interactive understanding of computing, that necessitates going far beyond Turing limit.

Introduction. Twilight of the Absolutes.
Meaning Makes the Difference

The world of omnipotent Turing-computable formal systems that could be used to reconstruct the Universe in its entirety proved to be yet another paradise from which we were expelled. Of historical absolutes nothing has remained today; no absolute time, space or vacuum, no preferred frame of reference. Earth is no longer the centre of the universe. We are becoming accustomed to the idea that the religion we are born to is only one of many in a global village. In short, there are no longer grounds for absolute truth.

The approach nowadays is increasingly pragmatic. We are not searching for absolute truth valid for the (one and only) Universe in general. We are searching for truthfulness – a reasonable and adequate approximation for the plurality of existing Universes - the best truth in given circumstances according to our best knowledge.

Through globalization, we are facing the question of multitude of contexts and we are only beginning to learn how to cope with the multitude of universes. Much can be learned from biological systems which through evolution have developed semantic metabolism as a cognitive response to the problem of shifting contexts.

Multi-context theories imply "local holism" which says that the meaning of linguistic expressions depends on local theory. The question is then how to define the rules for navigation across contexts and how to establish the identity of meaning of linguistic expressions from different theories.

Shifting the focus from semantics to pragmatics implies ascribing the central role to the *meaning* instead of the *truth*. Those two concepts however are inseparably entangled. It seems appropriate to talk about shifting the focus from (The) truth of a meaningful world to the meaningfulness of a truth(like) world.

Computers are information-processing devices that have changed dramatically compared to their original function which consisted in sequential processing of data. Contrary to traditional computation, in which computer provided with a suitable algorithm and an input was left alone to crunch the numbers until algorithm terminated, interactive computation (Goldin et al, 2005) implies interaction i.e. communication of the computing process with the external world during the computation. Computational processes are conceived as distributed, reactive, agent-based and concurrent.

Interaction consequently provides a new conceptualization of computational phenomena which involves communication and information exchange.

Background

Leibniz's dream of Mathesis Universalis, a universal science encompassing all existing knowledge, appears today to be a matter of the practical utilization of Informatics. The necessity of conceptualization of global informational space calls for an understanding across the borders of previously independent universes embedded in their local contexts. The construction of a universal knowledge system is clearly a considerably more complex task than was originally imagined and even the much more modest ambition of obtaining a smooth flow of knowledge between sub-fields of a multi-disciplinary area meets significant problems.

Each theory, no matter how formal, is embedded in at least two contexts: the linguistic context of natural language, and a situational context of the practical application.

Post-modernists deny that we can justify knowledge by reference to either empirical facts (pragmatics) or logical truths (semantics), because of the constructed nature of knowledge, so they endorse an “anything goes” philosophy. However, even recognizing the fact that *knowledge always is context-dependent*, it is possible to establish epistemology upon a practice (pragmatics) as a criterion of *meaningfulness* instead of searching for *absolute truths* in semantics.

Wittgenstein’s claim in Philosophical Investigations "Meaning just is use." presents possible grounds for a pragmatic approach to meaning that encompass language as both thought expression and speech act. It may also apply to information processing in physical systems such as living organisms. Acting in the physical world may be seen as a generalization of a language game in which linguistic symbols are replaced by physical objects such as e.g. molecules.

The problem of absolutes has become acute nowadays: no absolute time, space or vacuum, no preferred coordinate system; there is no longer firm ground for absolute truths. What remains however, is scientific *truthlikenss* - the best truth in given circumstances according to our best knowledge. There is an essential difference between truth and truthlikeness in that truth is absolute, objective and eternal, while truthlikeness is relative, constructed and evolving. The problem of

linguistic holism may be resolved by replacing identity with similitude and veracity with verisimilitude. We can learn from biological systems which through evolution have developed *semantic metabolism* (Maturana, Varela) as a cognitive response to the problem of shifting contexts.

Truth and Truthlikeness

Science is accepted as one of the principal sources of truth about the world. It might be instructive to see the view of truth from the scientific perspective. When do we expect to be able to label some information as "true"? Is it possible for a theory, a model or a simulation to be "true"?
Popper was the first prominent realist philosopher and scientist to adopt a radical fallibilism about science, defending at the same time the epistemic superiority of scientific method. Popper was the first philosopher to abandon the idea that science concerns truth and to take the problem of truthlikeness seriously. In his early work, The Logic of Scientific Discovery, Popper implied that *the only kind of progress an inquiry can make consists in falsification of theories.* (Popper, 1980)
Now how can a succession of falsehoods constitute epistemic progress? Epistemic optimism means that if some false hypotheses are closer to the truth than others, if truthlikeness (verisimilitude) admits of degrees, then the history of inquiry may turn out to be one of steady progress towards the goal of truth. (Oddie, 2001)
"While truth is the aim of inquiry, some falsehoods seem to realize this aim better than others. Some truths better realize the aim than other truths. And perhaps even some falsehoods realize the aim better than some truths do."
Kuipers (2000) developed a synthesis of a qualitative, structuralist theory of truth approximation:
"In this theory, three concepts and two intuitions play a crucial role. The concepts are confirmation, empirical progress, and (more) truthlikeness. The first intuition, the success intuition, amounts to the claim that empirical progress is, as a rule, functional for truth approximation, that is, an empirically more successful theory is, as a rule, more truthlike or closer to the truth, and vice versa. The second intuition, the I&C (idealization and concretization) intuition, is a kind of specification of the first."
According to Kuipers, the truth approximation is a two-sided affair amounting to achieving more true consequences and more correct

models, in a feedback loop, the position which obviously belongs to scientific practice. (Dodig-Crnkovic, 2004)

Search for Absolute Truth in Language through Formalization

The dream of a universal formal system that can be used to produce all truths and only truths within some area of knowledge is very old. Descartes' philosophy demanded that words in the scientific language should possess precise and unambiguous meanings. Leibniz developed an idea of a universal symbolic and logical calculus (calculus ratiocinator). The idea was to produce a completely rigorous and unambiguous language.

Leibniz hoped that a formal language would save us from the unnecessary ambiguity of the natural language. In the early 1920s, Hilbert's program for mathematics aimed at a formalization of all of mathematics in axiomatic form, together with a proof that this axiomatization is consistent. Whitehead and Russell's Principia Mathematica, the most famous work on the foundations of mathematics endeavored to deduce all the fundamental propositions of mathematics from a small number of logical premises, establishing mathematics as applied logic. However, Gödel, inspired by Hilbert's program, proved in 1931 that any such formalization is doomed to incompleteness.

Gödel's theorems (Gödel, 1992) show that in any sufficiently powerful logical system, statements can be formulated which can be neither proved nor disproved within the system, unless the system itself is inconsistent. Gödel's results are interpreted as the proof that there are limitations to the powers of any particular formal system. It is possible to re-phrase Gödel's argument in terms of text vs. context. Every formal system is surrounded by some context; it is never formulated in a vacuum.

Gödel's argument is often used to claim that strong artificial intelligence is impossible. Yet it has only been stated without any sort of proof that no such limitations apply to the human intellect (Dodig-Crnkovic, 2001). In what way then is Gödel's limit overcome in natural intelligence (natural language)? It's rather simple - natural language is both inconsistent and incomplete but – remarkably enough – it works!

The Ocean of Truth, the Islands of Theories

The minimum common structure in all languages appears to be logic. However, classical logic proves inadequate for the description of the entire real world. A simple logical structure is not even sufficient for the purposes of the complex world of science; hence the well-known paradoxes of physics such as the dual (particle-wave) nature of light.

In physics there are interfaces between different levels of abstraction (levels of common modeling language) in which separate adjacent universes of different scales must be connected by a type of translation mechanism, resembling a system of locks used to raise or lower boats from one water level to another. There is no formalism yet devised to derive a theory of human cell from first principles (axioms) with rules of inference. No one has even succeeded in deriving it from physics either. The similar is true for mathematics.

“You see, you have all of mathematical truth, this ocean of mathematical truth. And this ocean has islands. An island here, algebraic truths. An island there, arithmetic truths. An island here, the calculus. And these are different fields of mathematics where all the ideas are interconnected in ways that mathematicians love; they fall into nice, interconnected patterns. But what I've discovered is all this sea around the islands.”

Gregory Chaitin, an interview, September 2003

The ocean in Chaitin’s metaphor defines the context for all the different types of mathematical theories. A similar picture can be drawn for physics. The conventional approach is to assume that context as well as rule systems for sciences are fixed.

In adaptive intelligent behavior of agents such as individual humans, this might not be the case: neither context nor the principles (rules) are fixed. This gives flexibility to individual behavior that is advantageous from the evolutionary point of view. Of course, formal systems have a raison d'être of their own, in cases when, for the purpose of analysis, rules can be considered fixed, and the context unchangeable.

“The detailed study of the rules which work across contexts is exactly what is missing in Wittgenstein's approach, even if his philosophy clearly goes towards this clarification. This kind of study is also what is missing in the different attempts to face the problem of holism. All attempts to solve the problem of holism end up with a search of shared contents: communication is either the sharing of meanings or a convergence towards some shared meanings or contents. No question has been posed on the means to attain this aim; Davidson 1986 (p. 445) speaks of the "mysterious" aspect of the communicative success. On the contrary, the suggestion stemming from artificial intelligence is that there is no

mystery at all: we share and we may explicitly study general rules to navigate across contexts. For a communication to be successful, we need to share these high level rules, and the formal study of this kind of rules may help us to understand exactly the strategies used in successful communication." (Penco, 2001)

The Infinity of Language

Language semantics is a continuum in the sense of nuances and overlaps of meanings. A characteristic of a continuum is that it allows for the realization of infinity in a finite space. The world we live in is infinite. How do we cope with the infinity of information surrounding us?
An adult human brain has more than 10^{11} neurons which communicate through connections that form increasingly complex circuits (Damasio, 1999). Any particular neuron has between 10^4-10^5 links. The total number of connections in the human brain exceeds 10^{15}. The subtlety and complexity of the ways the neural network in our brains interconnect is amazing.
Moreover, each neuron has an astonishing number of built-in capabilities, its ability to conduct impulses (like a wire or an optical fiber) to attenuate signals (like a resistor), to integrate inputs (like a capacitor), to act as a power source (like a battery) and as a gate for thousands of other neurons.
"Hinton et al. (1993) conclude that the meaning of a word appears as if it were a point in a semantic space. The region around each word represents what in chaos theory is referred to as a point attractor. Once a neural network's state enters such a region, it will cause it to be inexorably drawn to the point represented by that word. Because such regions overlap, and because the semantic space is multidimensional, it becomes easy to see why an impaired system ends up in an adjacent region which has at its centre a point containing a word that looks like similar (a visual relative) or has a similar definition (a semantic relative). It must be obvious from this that the internal information environment comprises not only what information is stored by the recipient, but also how it is accessed and retrieved." Stonier, 1997
The complexity of our neural structure reflects the infinity of the universe that we are able to deal with, that is, visible in our language capability. Looking at the graphical representation of language such as Visual Thesaurus http://www.visualthesaurus.com/online it is obvious that

making detailed connections between the related words soon fills the entire space (semantic continuum).

The Semantic Metabolism

In trying to understand the meaning of meaning and truth and the role they play for semantics and pragmatics, it is useful to look back at ourselves as cognitive biological agents. A living organism can be fruitfully analyzed as an information processing system, or rather as a semantic metabolic system.

"The idea of semantic metabolism is this: when there is an information input into the human brain, such as a visual observation or an auditory message, the information is metabolized by the brain the way a molecule of glucose or an amino acid is metabolized by the cell, or the way a hormonal message entering the cell is cycled throughout the various cellular systems. (…)

Cells receive information from their environment all the time – information which is decoded by putting it into a chemical, metabolic or psychological **context**. Such a process can take place only because the cells provide an internal environment which allows them to respond to external chemical stimuli in a highly selective manner.", Stonier, 1997

Consequently, a biological system may be interpreted as an information system in which information stored in the DNA molecule is used to control the behavior of the cell. The meanings of different chemical structures consist very manifestly in their use. Applying Wittgenstein's vocabulary here, we can observe a "language game" in its primordial form. There is of course a symbolic counterpart used in mapping, describing and interpreting the processes taking part. But the "meaning" of strings of symbols is strictly their use in a given context.

Pragmatics - The Inevitability of Context

Pragmatics is the study of the ways that context affects meaning. The two primary forms of context important to pragmatics are linguistic context (i.e. the language surrounding the phrase in question) and situational context (i.e. every non-linguistic factor that affects the meaning of a phrase such as the people involved, the time, the location etc).

The question of traditional "objectivity" is the question of the possibility of the privileged (absolute) frame of reference. One of the consequences of

Einstein's relativity theory on philosophy is the abandonment of the idea of the absolute. What remains is a system of communicating frames of reference containing local universes with their local theories and local symbolic systems which exchange meaning.
[Postmodernists] "condemn the traditional ideal of objectivity not only as intellectually untenable, but also as inimical to freedom, and in its place they champion an 'anything goes' attitude to truth. (In addition to the works of Derrida, Foucault, and arguably Rorty, see J. Baudrillard, Simulations, trans. P. Foss, P. Patton, and P. Beitchman (New York: Semiotext[e], 1983).) They would have us abandon the very idea of objectivity. On the other hand, far too many opponents of post-modernism insist on a traditional ideal of objectivity as the only bulwark against an invidious culture of relativism and irrationalism, perhaps even social chaos. (A much discussed example is A. Bloom, The Closing of the American Mind, 1987. In many ways, however, the same might be said of J. Habermas, The Philosophical Discourse of Modernity, trans. F. Lawrence, 1987). They would have us ignore the manifest problems in the traditional concept of objectivity." Mark Bevir, 1999
Meaning is contextual with respect to language and the world, and it also actively affects other meanings and the world.

Interactivity and Logical Pluralism

Historically, science was forced to leave absolutes, one by one. We were shifted from the absolute center of the Universe with an unique and privileged coordinate system, and placed in the outskirts of our galaxy which in no way is special among galaxies, only to later on be forced to leave the idea of absolute space altogether and what is even worse to give up absolute time. Now it is time to leave the absolute truth, which is connected to leaving the idea of one and only true logic (logical monism).
How does the change in logic relate to computing, computers and information? Those elements influence each other and the development within one field induces the development in the others, which in its turn, influences the original field, and so on.
There are several points of departure one can take in order to explore the alternatives of logical monism in the context of Philosophy of Information and Computation.
Focusing on information instead of knowledge can be the smooth way to go from logical monism. The alternative, logical pluralism (Beall and Restall, 2000, 2005) is motivated by an analysis of disagreement within

the classical first-order logic, relevant logic and intuitionistic logic in the account of logical consequence (and hence of logical truth). Allo (2006) is arguing that logical pluralism could also entail semantic informational pluralism as informational content depends upon the underlying logic one assumes. Furthermore:

"An elementary consequence of this point of view is that, when a formal account of semantic information is elaborated, the absolute validity of logic cannot be taken for granted. Some further — external — evidence for its applicability is needed."

Allo presents an interesting, and for practical purposes relevant, case of communication between agents adhering to different logics in a multi-agent system. Taking examples from the Philosophy of Computing, I will illustrate why information pluralism (as a consequence of logical pluralism) is not only interesting theoretical problem, but has relevant practical consequences. Understanding of contexts where it appears may help us computationally articulate fields outside the domain of traditional computing.

This is the central point: information is something that is characteristic of a dynamical system; knowledge presupposes static, steady states. Knowledge is not something you receive today and discard tomorrow. Information is.

"I believe it inevitable that we revisit logic. Many have concluded this as well. (I've mentioned Barwise before.) Alternative logics already exist in fields that presently seem remote from science - in fact this is the point, they seem remote from science precisely because their logics are so different. I suggest we consider artistic and humanity-centric "logics" also, as we hunt for tools, and be open to a scope that includes internal conceptual mechanics: desires, intuitions, emotions, creativity." Goranson (2005)

The new interactive (communicative) role of computing is apparent in the Internet, the phenomenon that allows global communication and data transfer, making information easily available for people in different fields, establishing completely new preconditions for interdisciplinary learning, communication and collaboration. Related to the question of influence from other fields on computing, let us mention the work of Cantwell-Smith (1996).

In his book On the Origin of Objects, Cantwell Smith gives an outline of the foundations for Philosophy of Computing, which may be understood as a philosophy of the phenomena that produce, transfer, or preserve information. The book ascertains that the old digital, mechanical computing paradigm is not enough; there is only a vague intuition of something new that will result from the opening up of computing (as

defined by Hilbert's mathematical research agenda, i.e. algorithms) to the arts, humanities and other non-scientific practices. Let me illustrate by the following quotes:

"Not only are notions of mathematical proof being revised (...). Other distinctions are collapsing, such as those between and among theories, models, simulations, implementations and the like. " (p. 360)

"In the main the answer will emerge slowly, as appropriate vocabularies and intuitions are developed. But one thing can be said here. To the extent that the project is foundationalist or has foundationalist leanings on anyone's conception, it is intended to be a common foundation for everything, not just, more even preferentially, for the technical or scientific or "objective". (…) Hence the reference to CP Snow in the opening paragraph: the story is intended to be neutral in respect to – and thereby, perhaps, to help heal – the schism between the sciences and humanities." (p. 94)

Some years later, the positive side of what is going on become salient – computing is bringing together sciences and arts, in a development parallel to that of the Renaissance, (Dodig-Crnkovic, 2003), now with the computer in the place of the printing press:

"The important difference is that the computer (the physical object that is directly related to the theory) is not a focus of investigation (not even in the sense of being the cause of a certain algorithm proceeding in a certain way) but it is rather theory materialized, a tool always capable of changing in order to accommodate even more powerful theoretical concepts."

New technological developments are exposing new sides of our relations with each other, as articulated in the arts and humanities, as well as in our relations with nature, as expressed in sciences. These changes have of course feedback mechanisms. Technology changing culture in its turn changes technology.

What becomes especially visible is the *intentionality* of human actions, even the intentionality implicit in technologies. Computers are as much theoretical devices as the material ones. Our new aim is to make computers capable of accommodating natural computation, as the most expressive way of computation able to simulate natural phenomena, including cognition.

The possibility of choice and its consequences makes value systems one of central questions (Point (18) of Floridi's program). All this becomes the subject of the investigation of Philosophy of Information and Computing. Traditional computing is not enough; computing is expanding its domains. I definitely agree with the need for new logic, including logical pluralism. Actually pluralist logics are developing within the theory of computing

(Allo, 2006) and they will soon show as tools we need to re-conceptualize the world (or at least the computational theory of it). In terms of the new interaction paradigm computational processes are conceived as distributed, reactive, agent-based and concurrent. Agents, in general, may use different logics. Interaction provides a new conceptualization of computational phenomena which involves communication and information exchange, and makes way for logical pluralism.

Conclusions

One of the obstacles to the realization of Leibniz's vision of Mathesis Universalis is that all knowledge is context-dependent and always embedded in a natural language with all of its ambiguity. Attempts to automate knowledge discovery and communication between computerized knowledge bases meet the incompatibility of syntactically identical expressions with different semantic and pragmatic provenance. Coming from different universes, utterances with the same spelling may have a continuum of meanings – a problem that must be addressed.

The formalization question is related to the characteristics of the natural language semantic continuum. Through its evolution, the human brain has developed the capability to communicate via natural language. We need computers able to communicate in similar ways, which calls for a new and broader understanding far beyond the limits of formal axiomatic reasoning that characterize computing today, far beyond the Turing limit.

We are not searching for absolute truth or absolute certainty valid for the (one and only) Universe in general. We are searching a reasonable and adequate approximation for the plurality of existing Universes - the best truth in given circumstances according to our best knowledge and intentions.

Shifting the focus from semantics to pragmatics implies ascribing the central role to the *meaning* instead of the *truth*. Those two concepts however are inseparably entangled. It seems appropriate to talk about shifting the focus from (The) truth of a meaningful world to the meaningfulness of a truth(like) world. We are only beginning to learn how to cope with the multitude of universes. Much can be learned from biological systems which through evolution have developed semantic metabolism as a cognitive response to the problem of shifting contexts.

References

Bevir M., 1999: *The Logic of the History of Ideas*, Cambridge Univ. Press

Chaitin, Gregory J, 1987, *Algorithmic Information Theory*, Cambridge UP

Damasio A. R., 1999, *The Scientific American Book of the Brain*, New York: Scientific American

Davidson D., 1984, *Inquiries into truth and interpretation*, Oxford, Oxford University Press

Dodig-Crnkovic G., 2004, *Model Validation and Semantics of Information*, Proc. Model-Based Reasoning in Science and Engineering Abduction, Visualization, and Simulation, Pavia, Italy

Dodig-Crnkovic G., 2003, *Shifting the Paradigm of the Philosophy of Science: the Philosophy of Information and a New Renaissance*, Minds and Machines: Special Issue on the Philosophy of Information, http://www.idt.mdh.se/~gdc/work/shifting_paradigm_singlespace.pdf

Dodig-Crnkovic G., 2001, *What Ultimately Matters, Indeed?* Proc. Conf. for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, http://www.idt.mdh.se/~gdc/work/what_ultimately_matters.pdf

Gordana Dodig-Crnkovic, *Semantic of Information and Interactive Computation,* Logique & Analyse dedicated to "Logic and the Philosophy of Information"(December 2006 in Vol. 49 Issue 196) submitted http://www.idt.mdh.se/personal/gdc/work/SemanticsInformationInteractiveComputation.pdf

Gödel K., 1992, On Formally Undecidable Propositions Of Principia Mathematica And Related Systems, Dover Pubns; Reprint edition

Hinton, G. E., Plaut, D. C., and Shallice, T. (1993). Simulating brain damage. Scientific American, 269, 76-82.

Kuipers, T.A.F., 2000, From instrumentalism to constructive realism: on some relations between confirmation, empirical progress, and truth approximation - Dordrecht Kluwer Academic

Kuipers, T. A. F., 1987, ed., What is closer-to-the-truth? A parade of approaches to truthlikeness, Poznan Studies in the Philosophy of the Sciences and the Humanities, Volume 10, Amsterdam: Rodopi

Kuipers, T. A. F., 2002, Inductive Aspects of Confirmation, Information, and Content, To appear in the Schilpp-volume The Philosophy of Jaakko Hintikka

Kuipers, T. A. F., (to appear), Empirical and conceptual idealization and concretization. The case of truth approximation, in (English and Polish editions of) Liber Amicorum for Leszek NowakMalpas J., 2003, *Donald Davidson*, Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/davidson

Oddie, G., 2001, Truthlikeness, The Stanford Encyclopedia of Philosophy, E.N. Zalta (ed.), http://plato.stanford.edu/archives/fall2001/entries/truthlikeness/

Penco C., 2001 Local Holism" in V. Akman, P. Bouquet, R. Thomason, R.A. Young (eds.) Modeling and Using Context - Proceedings of the Third International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'01) , [Lecture Notes in AI Volume 2116] Springer Verlag (pp.290-303)

Popper, K., 1980, *The Logic of Scientific Discovery*, Unwin Hyman Ltd, London. Translation of Logik der Forschung

Quine, W.V.O., 1964, *Word and Object*, The MIT Press, Cambridge

Russell B., 1921, *The Analysis of Mind*, Essays http://www.literaturepage.com/read/russell-analysis-of-mind.html

Stonier, T., 1997, Information and Meaning. An Evolutionary Perspective, Springer, Berlin, New York

Wittgenstein, L., 2001, Philosophical Investigations. The German text, with a revised English translation, 3rd ed., Blackwell Publishing

Wolfram, S., 2002, A New Kind of Science, Wolfram Media, Inc.
Wright, R, 1988, Edward Fredkin in .Three Scientists and Their Gods: Looking for Meaning in an Age of Information, Times Books

COMPUTATIONAL MODELING OF AUTONOMOUS MACHINE THINKING

JUYANG WENG
Department of Computer Science and Engineering

Michigan State University

East Lansing, MI 48824 USA

weng@cse.msu.edu

Abstract

Keywords: Mental architecture, AI architecture, computational neural science, autonomous agents, cognitive development, perceptual development, behavior development.

As early as the 1950s, in his article .Computing Machinery and Intelligence. [1], Allen Turing raised a fascinating question: .Can machines think?. However, Turing avoided a de_nition of thinking. As a variant of his original question, Turing suggested in that article to consider instead his .imitation game,. now known as the Turing test. He did not regard his imitation game as equivalent to the original problem of machine thinking. He wrote:

.We cannot altogether abandon the original form of the problem. I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted..

This period did not come as early as Turing predicted. As the page of the 20 century was turned, humans had gained a better understanding about what computers have done well and what they have not. Consequently, the important question about machine thinking has somehow withdrawn out of the spotlight of attention. Many computer scientists do not believe that a machine can think in the way humans do. The main reason for this situation is that there is still not a clear understanding of machine thinking. Machine computations

in the traditional non-developmental paradigm are too far from human autonomous thinking activities. After reviewing some existing studies about thinking, this paper puts major missing architecture pieces together to show the contrary . yes, machine thinking is much harder than people once thought, but a developmental robot can think. We should not expect that human-like thinking is built into a developmental robot quickly overnight. Like a human, the skills of thinking are gradually developed through interactions with the physical world. They develop in parallel with other perceptual, cognitive, behavioral and emotional capabilities.

As one may expect, this paper cannot explain *completely* how the thinking process works in humans, since human thinking involves some components that are not fully understood at the current state of knowledge. Inspired by biological thinking process, this paper discusses basic mechanisms[1] that are essential to understanding how the scaffolding of autonomous thinking takes place in humans and robots computationally, and to enabling autonomous thinking in a developmental robot.

One cannot clearly explain what autonomous thinking is without describing a mental architecture that enables autonomous thinking. Therefore, this paper _rst describes a biologically motivated computational model of mental architecture. Then, it explains autonomous thinking as autonomous mental processes that are supported by skills of perception, cognition, behavioral and motivation (which includes emotion) which are incrementally developed.

After introducing some simple mental architecture types, as shown illustrated in Fig. 1, the paper introduces Type-4 mental architecture:

De_nition 1 (Type-4) *The* Type-4 mental architecture *is a Type-3 mental architecture, but additionally, the internal voluntary decision is sensed by the internal sensors* S_i *and the sensed signals are fed into the entry point of sensors, i.e., the entry point of the attention selector* T*. In order to recall the effects of the voluntary actions, not only the expected reward value is estimated by the value system, but also the primed context which includes the primed action and the primed sensation.*

The developmental of autonomous thinking starts from

simple re_exive sensorimotor skills in a developmental
agent. While the sensorimotor skills are developed incrementally
and become more and more skillful, they can handle
more and more complex tasks. Thinking does not have

[1]Not exhaustive at all.

t
Time
S_{e2}
S_{e1}
E_{e1}
E_{e2}
S_i
$S_{i\,3}$
Sensors:
Effectors:
Last
context
Primed
context
E_{i1}
R M
D
V
T
$_2\,S_{i\,1}\,E_{i2}$
L

Figure 1. Progressive additions of architecture components from Type-2 to Type-5. Type-2: adding attention selector T and its (internal) control input E_{i1}. Type-3: Adding motor mapping M and its (internal) control E_{i2}. Type-4: Adding internal controls S_{i1} and S_{i2} and the primed sensation S_{i3} to the entry port of perception T. The block marked with D is a delay module, which introduces a unit-time delay for the corresponding vector. Type-5: Developmental T, R,M and V .

to be totally covert, in that no external behaviors are displayed during a thinking process. In fact, humans often perform tasks while thinking. The process of thinking can either involve external behaviors or display little or no external behaviors. The effect of thinking can be display either immediately or later.

Building on the architectures de_ned, the paper introduces the concepts of external and internal reasoning process.

De_nition 2 (External and internal reasoning process) *There are three types of reasoning processes, external, internal, and mixed, corresponding to the attention in which the attention module* T *attends to external, internal or both, respectively.*

Next, the paper establishes that the architecture Type-4 enables internal reasoning, modeled as a complex type of sensorimotor skills.

Theorem 1 *The Type-4 architecture allows internal reasoning to realize the following kinds of learning (1) nonassociative*

learning, (2) classical conditioning, and (3) instrumental conditioning.
A proof will be provided in the full paper.
The next theorem establishes that autonomous planning is also enabled by such an architecture.
Theorem 2 *The Type-4 architecture allows internal reasoning to realize autonomous planning.*
A proof will be provided in the full paper.
The paper further introduces architectures Types 5 and 6 that are more powerful for the development of thinking capabilities.
Next, this paper discusses our experimental results which showed how early learned simpler sensorimotor skills can be transferred to new and more complex task settings, a process that demonstrates how simpler mental skills can be developed into more complex ones.
In this experiment, the process of .arranged experience. is in line with the concept of .zone of proximal development . (ZPD) proposed by Lev Vygotsky [2]. ZPD is a latent learning gap between what a child can do on his or her own and what can be done with the help of a teacher. Wood, Burner & Ross [3] used the term .scaffolding. to describe such an instructional support through which a child can extend or construct current skills to higher levels of competence. When the children become more skillful, the scaffolding (arranged experience) is slowly removed.

7. References

[1] A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433.460, October 1950.
[2] L. S. Vygotsky. *Thought and language*. MIT Press, Cambridge, Massachussetts, 1962. trans. E. Hanfmann & G. Vakar.
[3] D. J. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry*, pages 89.100, 1976.

Juyang (John)Weng received his PhD degree in Computer Science from University of Illinois, Urbana, IL USA, January 1989. He is a professor at the Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA. His research interests include computer vision, speech recognition, human-machine multimodal interface using vision, audition, speech, gesture

and actions, and intelligent robots. He is the author of over one hundred research articles and book chapters. He is a coauthor (with T. S. Huang and N. Ahuja) of the book *Motion and Structure from Image Sequences*
2
(Springer-Verlag, 1993). He is an editor-in-chief of *International Journal of Humanoid Robotics*, the chairman of the *Autonomous Mental Development Technical Committee of the IEEE Neural Networks Society* and the chairman of the Governing Board of International Conferences on Development and Learning. He was an associate editor of *IEEE Trans. on Pattern Recognition and Machine Intelligence* (2001-2004) and *IEEE Trans. on Image Processing* (1994-1997), a program co-chair of the *NSF/DARPA Workshop on Development and Learning* (WDL), held April, 5-7, 2000 at Michigan State University, East Lansing, MI (http://www.cse.msu.edu/dl/), and a program co-chair of the *IEEE 2nd International Conference on Development and Learning* (ICDL'02), held at Cambridge, MA, June 12-15, 2002 (http://www.egr.msu.edu/icdl02/). His home page: //www.cse.msu.edu/ weng/.
3

EVERY METAPHOR HAS ITS "PLACE"

WHY SIMILAR EMBODIMENT IS NECESSARY FOR HUMAN-AI COMMUNICATION

ROBIN L. ZEBROWSKI

Department of Philosophy

University of Oregon

(541) 346-5550. rzebrows@uoregon.edu

Traditionally, artificial intelligence has ignored the role of the body in thought and cognition. Although there are an increasing number of researchers engaging in embodied AI, many of them still do not acknowledge most of the theoretical issues underlying the importance of embodiment. Furthermore, almost every one of those individuals working in embodied AI is working on a very low level without reference to the complex issues in cognition and communication that have driven AI from the beginning. However, there is good evidence to show that the composition of our bodies actually structures our conceptual abilities and understandings, often through metaphorical conceptualization. Many of our abstract concepts have been built off of our bodily interactions with the world, and hence they structure how we are able to think about things and what things we are able to think about. There is both empirical evidence for this and a robust evolutionary story behind it. If this is true, any artificial intelligence we may someday build will necessarily have to share some very specific sorts of embodiment with us for us to ever recognize such a creature as intelligent, and also to make possible any sort of meaningful communication between us.

"Unless I have an exterior others have no interior"(Merleau-Ponty, 373)

It is hard to deny, from any perspective, that artificial intelligence is a struggling field. It has promise, sure, and it moves forward, in small steps, but Alan Turing once predicted that we would be living amongst thinking machines by now, and we are hardly closer to that reality than we were when the prediction was written over fifty years ago. AI has proceeded from a generally materialist stance, but this has manifested in modern AI with an unmistakable dualism. That thought might be computation is not what ought to be debated, but instead that it might be pure computation in the void – disembodied, as it were, or rather, embodied in any sort of system whatsoever – this is the problem. Rather than traditional researchers who attempt to build abstract symbol systems that we hope will somehow ground meaning and gain intentionality, there are some theorists who rightly argue for necessarily embodied AI: real physical systems in the real world. This is the correct place to start, but it alone is not enough. As this paper argues, it is required that AI be not only embodied, but embodied in some ways similar to ourselves if we hope to successfully communicate with it (and hence to recognize it as intelligent at all).

For the last twenty-five years, cognitive linguistics has offered us a robust theory of embodied metaphor. First argued by George Lakoff and Mark Johnson in 1980, this theory argues that much of our language is heavily reliant upon our specific type of embodiment in our specific type of environment. The body remains a constant factor throughout in how we structure our metaphors to ground our concepts in the world. Sure, machines are highly sophisticated now and have tremendous memory and search capabilities, and our algorithms and neural nets are capable of apparently intelligent performances. But these mean nothing in the absence of the environment, and furthermore they mean nothing in the absence of a body.

However, it is not just any body that is needed for AI-human communication. For example, often considered the most disembodied and abstract of our concepts, many of our logical abstractions and understandings come directly from our sort of embodiment in the world. For instance, we see conclusions as *resting upon* premises. We talk about our abilities to *grasp* a concept. We can talk about the *foundation* of an argument, or how it is on *shaky ground*. These terms are not

computation in the void – they come from our physical interactions with our specific environments through *our* sorts of bodies. The body itself is the relevant inference-making mechanism when it comes to conceptualization. As Lakoff and Johnson put it, "a metaphor can serve as a vehicle for understanding a concept only by virtue of its experiential basis" (Lakoff and Johnson 1980,18). While our bodies do differ in some important ways, an idea which deserves much more attention than it has been given, we share enough of *something* to ground our interaction and communication, and AI must also share this something in order for us to even recognize it as intelligent, should that day ever arrive. Our bodies were built through interaction with this world, and the cultural aspects of our world, as well as much of the physical environment we find ourselves in, were built through interaction with our bodies. Until we find a way for AI to participate actively in these worlds rather than hoping for some stroke of luck that bestows meaning onto the system mysteriously, AI will continue to fail. AI absolutely can succeed, but the field must take a broader view on what it means to think and be intelligent, and recognize that human communication goes way beyond the transmission of pre-packaged bundles of symbols (abstract thought) thrown back and forth via language. Examination of the body, and how it may metaphorically help us structure the world, is the only way that this will happen.

First, let us take a very brief look at what AI has traditionally said in relation to the body, and how this has affected the theories of conceptualization that have dominated the field. Then, let us look at the revolution that has been occurring within cognitive science in relation to embodiment, and how this feeds the theories of metaphoric conceptualization that I am claiming are necessary for AI to move forward. Then, I will enter very briefly into the thorny aspect of this argument: the fact that there is no real standard body amongst us. Lastly, I will make a case for why, without this particular type of conceptual semantics proposed by Lakoff and Johnson, we will never be able to communicate with any sort of AI we might build, and hence, never be able to recognize it as intelligent at all.

Traditional AI

The traditional view of cognition within AI is that of the simple symbol system. This view of traditional AI was best described in 1976 by Newell and Simon in their ground-breaking paper, “Computer Science as Empirical Inquiry”. Newell and Simon claimed that, “symbols lie at the root of intelligent action,” and went on to advocate what they called the

Physical Symbol System Hypothesis (Newell and Simon, 113-126). As they put it quite simply, “a physical symbol system has the necessary and sufficient means for general intelligent action.” They define intelligent action as “the same scope of intelligence as we see in human action” (Newell and Simon, 116).

Unfortunately, for many years in AI, no one took up this challenge to recreate human intelligence as human *action*. Instead, it was simply a case of trying to recreate human language, with the assumption that carrying on a conversation or displaying information about the world (without the machine ever interacting seriously, if at all, with that world) was enough to demonstrate whether or not intelligence had been achieved. "Human action" was replaced with "human conversation" and AI proceeded from there. In fact, many of the philosophical theories of mind that informed the strongest and most promising forms of AI were built on something like this entirely disembodied standpoint. So while the claim remained that thought was realizable in many physical systems and thus not some sort of abstract dualistic substance, it manifested itself in the field as pure symbol manipulation in the void, which, really, is dualism itself.

There is no great consensus on the formation of concepts and categories, or even what these terms mean necessarily. However, traditionally, the analytic school that most often informs the theories and foundations of AI has held concepts to be a sort of attribute list that picks out objects in the world by providing necessary and sufficient conditions for inclusion in that concept. While it is clear that many years of additional psychological work and philosophical examination of that work have called this theory of concepts into question, it remains one of two main conceptual theories within AI, the other being modal epistemic logic (Thomason, 2005). A look at any one of many traditional AI programs written in the last 40 years will show this conceptual scheme in play, almost certainly.

Keep in mind that functionalism, a view proposed back in the early days of AI, claimed that any material *in any form* could take up the task of intelligence and consciousness since it wasn't form but function that ought to be stressed as the vital component. As long as we could demonstrate the proper relationship between structure and logical states of an algorithm, then we were effectively demonstrating the same relationship between physical and mental states of an individual, and hence thought ought to arise as much in another substance as in ourselves (Putnam 1975, 291-303). This seems to be the view still shared in most of the AI community overall.

In an interview somewhat late in his career, Hilary Putnam, a name once synonymous with functionalism, nicely summarizes both the motivation behind the very functionalism that still drives much of the research in AI, and also his own revision of the theory including, most importantly, what he still believes is relevant as far as bodies are concerned. He says:

> Functionalism was itself a reaction against the idea that our matter is more important than our function, and that our *substance* is more important than our *activity*. My functionalism argued that, in principle, a machine (say, one of Isaac Asimov's robots), a human being, a creature with a silicon chemistry, and a being with an ectoplasmic body could all work much the same way when described at the relevant level of abstraction, and they would all be conscious, all have feelings, etc. That much I have not given up (Putnam 1999, 47).

Putnam is right about function being more important than substance: if anyone believes substance matters more, they certainly are not engaged in work in artificial intelligence. However, Putnam's revision, wherein he argues that even a being with an "ectoplasmic body" would be conscious and have feelings, is what is relevant and interesting here. What sorts of concepts would we share, or could we share, with a creature whose body is that different than our own? It is a purely fanciful question, but one which deserves at least some exploration.

A creature with an ectoplasmic body would not share our comportment toward the world in many important ways. Lacking a skeleton like our own, it would possibly have a different center of gravity. It would almost certainly lack the solidity of our own bodies, and so would not share our abilities to hold things, stand solidly upon things, and hence, completely lack our understanding of the solidity of both the world and our own bodies in and against that world. Would the metaphorical concept of a *solid foundation* in an argument exist for such a creature? Could such a creature ever fully appreciate the metaphorical *grasping* of an idea, and the importance of such a grasping? It seems as if we can point out numerous places where our bodies would differ enough to create fundamental schisms in our concepts, leaving communication extremely difficult. So, while Putnam's concentration on function over substance was a brilliant move that brought AI forward in spectacular new ways, there is a definite limit on the analogy that has since gone unrecognized. Please note, however, that this criticism does not demand we give up either consciousness or feelings in that ectoplasmic body; only that we lose every hope of ever communicating with it. An octopus might be a nice real world counterpart – a creature that humans have finally come to

accept has intelligence, but one so mysterious to us as to be virtually unreachable.

Historical Embodiment

However, after years of people in AI trying to work on mind-software totally independently of any body-hardware, there are now some researchers who do take seriously the argument that we need to pay real attention to the body in the real world. For the most obvious example, we need only look to Rodney Brooks, who first started arguing at least as early as 1987 that embodiment was a vital piece of the AI puzzle. In "Intelligence Without Representation," he tried to convince people working in the field that bodies were a vital, indispensable aspect of intelligence, and he did so from a purely engineering standpoint, recognizing that there would be philosophical implications but not proceeding from them. He not only argued that "at each step we should build complete intelligent systems that we let loose in the real world with real sensing and real action," but he included in this argument the idea that "mobility, acute vision and the ability to carry out survival related tasks in a dynamic environment provide a necessary basis for the development of true intelligence" (Brooks 1999, 80-81). Brooks has called for a type of research that demands a very specific sort of embodiment, and while he does not go so far as to analogize the mobility, vision, and survival-behavior with our own sorts of bodies, it does not seem unreasonable to examine, or at least note, the overlap of these traits with how our own bodies function. Many of the metaphors selected by Lakoff and Johnson throughout their work show a strong preference for mobility and vision, perhaps because evolution selected for these as among our most vital tools for survival and reproduction. If we take an evolutionary view of mind, it makes sense that those things most vital for our survival play large roles in our conceptual understanding of the world. Many philosophers both before and after Brooks have argued something similar – that the body is more than just a container for the mind, an incidental shell that takes whatever shape it needs to and holds the same stock of concepts. One hundred years ago Pragmatists like John Dewey made this argument, and then fifty years after that phenomenologists like Maurice Merleau-Ponty made it again. Recently, cognitive scientists like Andy Clark have continued in the same vein. Each time it is revisited, it is done so through a unique starting point, adding to the overall picture that we have of embodiment today. In more recent years, neuroscientists are joining the philosophers to weigh in on the issue and providing

valuable evidence that the body is not something separate from the brain – the two are one and the same system, and furthermore that language is not equivalent to thought. As Antonio Damasio puts it, "I am not saying that the mind is in the body. I am saying that the body contributes more than life support and modulatory effects to the brain. It contributes a *content* that is part and parcel of the workings of the normal mind"(Damasio, 226).

This information doesn't condemn AI, but the field in general does need to stop ignoring or denying it in favor of 50-year old non-productive theories and begin to take it seriously. Of the few researchers actively engaging in robotics for this purpose, almost none are paying attention to the humanoid aspects – which are as important as embodiment itself - if we ever hope to communicate with these creatures. Rodney Brooks has addressed some of these failings as well, concluding that perhaps we're "actually missing something in our models of biology… some 'new stuff' that we need" in order to achieve success in this field (Brooks 2002, 184). It seems to me not so much some "new stuff" that we need, as a new way of viewing what we already have – a body in the world which is intricately related to our brains and hence, minds; a body which structures our concepts, necessarily.

In the books *Metaphors We Live By* and *Philosophy in the Flesh*, Mark Johnson and George Lakoff argue that it isn't simply having *a* body, but having our *types* of bodies in our specific environment that provide us with the raw materials of thought. They argue that "the very properties of concepts are created as a result of the way the brain and body are structured and the way they function in interpersonal relations and in the physical world" (Lakoff and Johnson 1999, 37). To begin with, to claim that our language capabilities arise spontaneously in evolution and are unique to humans would seem an odd move, and the theory of conceptual metaphor they put forth recognizes that our language builds itself off "lower" structures present in other life forms – specifically the sensorimotor system. Since Lakoff and Johnson argue that conceptual metaphors are "mappings across conceptual domains that structure our reasoning, our experience, and our everyday language," it seems hard to imagine an AI succeeding without a similar conceptual system (Lakoff and Johnson 1999, 47). And even if we were able to build a machine with another sort of conceptual system, it seems unlikely that enough of our concepts would overlap with it to enable any sort of useful communication. We can simply look to the rest of the animal kingdom to see that humans are slow to acknowledge intelligence when it does not conform to a sort immediately recognizable to us. The large number of people who still refuse to accept animal intelligence of any sort provide as

much evidence that we need shared concepts for communication as any other argument I might offer in AI. And while Lakoff and Johnson seem doubtful on the topic of AI, they argue that these metaphors are built up and persist via neural selection once the body begins to form the connections (Lakoff and Johnson 1999, 57). This seems consistent with AI in general, both with connectionism and the more embodied robotics, so it should be considered a possible research route that has yet gone largely unexplored.

To give an example of this notion of conceptual metaphor, take the most disembodied activity we can often think of – logical reasoning. Johnson begins by simply showing that the language of bodily experience pervades our discussion of reasoning. For example, Johnson says of formal reasoning:

> When we reason, we understand ourselves as starting from some point (a proposition or set of premises), from which we proceed in a series of steps to a conclusion (a goal, or stopping point). Metaphorically, we understand the process of reasoning as a form of motion along a path – propositions are the locations (or bounded areas) that we start out from, proceed through, and wind up at. Holding a proposition is understood metaphorically as being located at that point (or in that area) (Johnson, 38).

And our language about reasoning reflects this metaphor. Even our logical operators are often arrows, forcing us to move (metaphorically and literally) to a new location! Johnson goes on to admit:

> We can, and do, extract away from this experiential basis, so that sometimes it looks as if we are operating only with a priori structures of pure reasoning; however, the extent to which we are able to make sense of these extremely abstract structures is the extent to which we can relate them to such schematic structures as connect up our meaningful experiences (Johnson, 64).

Furthermore, there is experimental evidence within neuroscience to back up these claims. Tim Rohrer performed fMRI and ERP studies to discover if the somatotopic maps of sensorimotor cortices that are well-established and accepted would be activated in both the literal and metaphorical cases. His claim for the importance of this work is that it "suggests that there is a functional contribution from the motor and somatotopic neural maps to semantic processing – contra modularist arguments that suggest that language is neurally encapsulated and borrows no or minimal structure from perceptual processing" (Rohrer, 1). He also stresses that

metaphor ceases to be the traditional "mere matter of language" and instead becomes "the result of *systematic* patterns of conceptualization" (Rohrer, 2). The results of his work showed that the metaphoric use of bodily terms activated the somatosensory maps much the same as they were activated for regular use of those terms, implying that our bodies are very actively involved in our use of even metaphorical terms.

Standard Bodies?

The biggest hurdle to be overcome, metaphorically speaking, is the mistaken belief that in arguing for similar embodiment I am actually advocating for a sort of standard body. Let it be clear that I am not. There is an abundance of evidence that there truly is no such thing as a standard body. We can look to neuroscience, biology, phenomenology, social philosophies, and disability studies, all of which provide some amount of evidence that human beings, although sharing a physical environment and *some* sort of embodiment, do not adhere to some sort of standard body that we can hold up as the ideal body that transcends cultures and individuals. Let us briefly look at this evidence in order to dispel any possible misunderstanding.

First, it should be noted that Lakoff and Johnson do, in fact, argue that we share a type of embodiment and that it is this very shared embodiment that causes some universality amongst our concepts. While I do believe that it is their research on metaphor that needs to be acknowledged within AI, I disagree about the universality of our embodiment and the implications this has for shared communication. They claim that "when the embodied experiences in the world are universal, then the corresponding primary metaphors are universally acquired" (Lakoff and Johnson 1999, 56). My claim, however, is that while there is no universal body, there is *something* shared amongst us, and that something is bodily based, enabling both our conceptual metaphors and our ability to communicate with one another. Let us briefly examine the arguments against a standard body before moving on to discuss just what is at stake in communication.

The strongest arguments for a standard body are likely to come from biology, so we can look first to biology to show us why this would be in error. Each human body is non-trivially unique. As Armand Marie Leroi reminds us in his book *Mutants*, "Each new embryo has about 100 mutations that its parents did not have." And "Not only are we each burdened with our own unique suite of harmful mutations, we also have to cope with those we inherited from our parents, and they from theirs,

and so on" (Leroi, 18). There is no single standard genome – we share most of our genes with other animals, and differ greatly amongst our human counterparts.

In a similar way we can see that in spite of the fact that psychologists have long spoken of "the brain" as if there were only one, in reality brains differ quite a bit from one another. More importantly, however, is the fact that biology does not dictate precisely how the body and brain will develop; rather, a good deal of environmental input is required for brains to develop in the way that the statistically largest number of brains develop. There are critical periods of development in which certain levels of, for example, sound or visual input are required for the brain and body to develop along pathways considered "normal." A good deal of active molding is required before brains and bodies will develop in "standard" ways. It seems, then, that since so much environmental input is actively required, we would be mistaken to call this standard, in any way other than a statistical average.

Phenomenology and social philosophies have provided additional support for the idea that there is no standard or universal body. To begin with, there is no experience of a standard body. For definitional purposes, we must distinguish between a standard body, which seems to contain some sort of "ought" in its definition (what one ought to look like or how one's body ought to be structured) versus a definitional average, which, although undeniably real, need not be a body that actually ever appears in the population. It is the body with the "ought" that does not actually exist, at least outside of a given cultural milieu, in spite of a strong intuition otherwise. Even within phenomenology, we can find divergence in what sorts of bodily experiences are reported by different people. Maurice Merleau-Ponty spent his career elaborating on a phenomenology of bodily experience and how that experience structures who we are and how we exist in the world. And yet even his research on the experience of bodies has been taken to task by other philosophers who point out that standardizing bodies across time, gender, social class, and economic status is an impossible task (Young, 1990). While it would be hasty to point only to experience as evidence of there not being a standard body, combined with the previous claims about biology and neuroscience we can only be left wondering whose body would be the standard, should such a thing exist. We are all embodied, and this, I argue, is what makes communication possible. Yet every body is non-trivially distinct, and perhaps it is this trait which makes communication interesting.

Communication

The real lesson to take from this argument is not simply that traditional AI has been and will continue to be a failure in its current, disembodied, dualistic form. It is also not that embodiment is the necessary step we must take in order to succeed, since several roboticists and researchers have made similar claims. It is, in fact, to show that without embodiment *very much like our own*, we will ultimately fail to communicate with intelligence, should we manage to create it. While the definition of intelligence varies with every philosopher and theorist who uses it, most of AI ought to still have some relation to Newell and Simon's original claim – intelligence is recognizable as human action, and our concepts come directly from this action and ought to reflect it accordingly. The notion of the Turing Test has rightly fallen out of favor, being exposed for the poor indicator of real intelligence (or consciousness, or mind) that it would be. We can safely assume computing power will solve the Turing Test well before an actual intelligent machine does. Furthermore, none of our language would have meaning for a creature without our spatial orientation toward the world or our bodily aspects to interact with that world. No creature without hands or limbs in some important way similar to our own will ever communicate the concept of "grasping" ideas, or perhaps even understand an idea as the sort of thing which can be grasped at all, in the way we understand it. Rationality and reasoning, the very pillars of the AI research project, are at stake when we talk about humanoid embodiment. This rationality begins right away upon entering this world. As Mark Johnson puts it, "Image schemas […] are the earliest and most primitive bases for the infant's emerging sense of the world. Such structures of bodily experience are nonpropositional and arise prior to language, and yet they are what make it possible for the child to make some sense of its surroundings and to act intelligently to achieve its ends" (Johnson 1999, 99). Our bodies are part of the concept-forming equation in an indispensable way. Without these specific types of bodies in this environment that built them, our concepts on their most basic level would be entirely different, in structure and content, although they would still most likely exist in some form unrecognizable to our current selves. Since meaning arises from these structures, AI is faced with a very difficult task in this respect – it is neither cheap nor easy to engage in a project of such magnitude as I claim is necessary for the success of the field. However, without accounting for how humans bodily acquire semantic content in this way, AI will just continue to produce expert systems at best, with no hopes of achieving true intelligence, in the sense

of independent thought and consciousness. When dealing with something as important as AI, we need to tread lightly, and to do so, we must keep a set of feet beneath us.

References:

Brooks, Rodney. *Flesh and Machines*, Pantheon Books (2002).

Brooks, Rodney. *Cambrian Intelligence: The Early History of the New AI*, MIT Press (1999).

Damasio, Antonio. *Descartes' Error*, Harper Collins (1994).

Johnson, Mark. *The Body in the Mind*, The University of Chicago Press (1987).

Johnson, Mark "Embodied Reason" in *Perspectives on Embodiment: The Intersections of Nature and Culture*, (Gail Weiss, ed.) Routledge (1999)

Lakoff, George and Mark Johnson. *Metaphors We Live By*, University of Chicago Press (1980).

Lakoff, George and Mark Johnson, *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*, Basic Books (1999).

Leroi, Armand Marie. *Mutants: On Genetic Variety and the Human Body*, Viking Press (2003).

Merleau-Ponty, Maurice. *Phenomenology of Perception*, Routledge Press (1962).

Newell, Allen and Herbert Simon. "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM* 19, no. 3 (1976).

Putnam, Hilary. "Philosophy and Our Mental Life", *Mind, Language, and Reality*, Cambridge University Press (1975).

Putnam, Hilary. *Key Philosophers in Conversation: The Cogito Interviews*, Routledge Press. (1999).

Rohrer, Tim. "Understanding Through the Body: fMRI and ERP Investigations into the Neurophysiology of Cognitive Semantics" (2005).

Thomason, Richmond. "Logic and Artificial Intelligence", The Stanford Encyclopedia of Philosophy (Summer 2005 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2005/entries/logic-ai/>; < http://plato.stanford.edu/entries/logic-ai/#knowledge>.

Young, Iris Marion. *Throwing Like a Girl and Other Essays in Feminist Philosophy and Social Theory,* Indiana University Press, (1990).

BOO-HOORAY AND AFFECTIVE APPROACHES TO ETHICAL TEXTUAL ANALYSIS

Carson Reynolds
Department. of Information Physics and Computing
University of Tokyo
+81 3-5841-6937
carson@k2.t.u-tokyo.ac.jp

In Language, Truth, and Logic Ayer took the surprising position that ethical judgments, exhortations, and descriptions are neither true nor false, but are emotive statements. While his argumentation was in support of the logical positivist position, it suggests a novel method for computational systems to recognize utterances regarding ethics. Namely, expressions of emotion can be likened to "boos" or "hoorays" issued from a metaphorical crowd in moral response. This paper presents a simple natural language processing system that searches for terms and categorizes the text accompanying these terms as a "boo" or "hooray," making use of a list of emotional terms compiled by Cowie et al. and orientations recorded by Whissel and Plutchik. The system uses this bag-of-words and a search engine to assign emotive scores to terms of the user's choosing. The contribution of this work is a primitive technique for computers to ethically evaluate textual queries.

Keywords: Computer Ethics, Affective Computing, Natural Language Processing, Emotivism, Emotion Recognition

COMPUTATIONAL EMOTIVISM

Emotivism, is a position that “judgments of value” are statements having neither truth nor falsity. More complexly, Ayer argues in Chapter 6 of Language, Truth, and Logic that a variety of statements regarding ethics cannot be the domain of philosophy [1952]. He describes four classes:

1. “propositions which express definitions of ethical terms, or judgments about the legitimacy or possibility of certain definitions”
2. “propositions describing the phenomena of moral experience, and their causes”
3. “exhortations to moral virtue”
4. “actual ethical judgments”

Of these, Ayer contends that only the first is “ethical philosophy.” The remainder he argues can be dealt with by the social sciences or classed as “mere pseudo-concepts.” In his view we “cannot argue ... the validity of these moral principles. We merely praise or condemn them in light of our own feelings.”

Recent research in affective computing [Picard, 1997] has pointed at a variety of methods for using computers to recognize information dealing with feelings. For instance, Liu et al. have demonstrated a system that analyzes text using common-sense reasoning to assess the (positive or negative) valence of text [2002].

In light of this recent work on computational processing of information dealing with feelings, a sort of “Computational Emotivism” is possible in which statements are analyzed for their content to determine if writers referring to the statement are “praising” or “condemning.” Or, more colloquially, if a group of documents is shouting “boo” or cheering “hooray” with regard to a statement.

RELATED APPROACHES

A variety of different approaches have been considered for trying to make computers behave in an ethical manner. These include Weld and Etzioni's work to include the notion of "harm" into a planner to create ethical "softbots" [Weld, 1994]. Eichmann proposed an ethic for Internet agents and spiders to limit bandwidth abuse [Eichmann, 1994]. More germanely, LaMuth described an expert system with “affective language analyzer” that guides “motivational determination” of an artificial

intelligence system [2003]. Additionally, there has been much work regarding textual analysis of affect that differs in approach, but often seeks the common goal of recognizing the sentiment or affective content of text. For instance, Pang et al. applied scores related to positive or negative sentiment [2002].

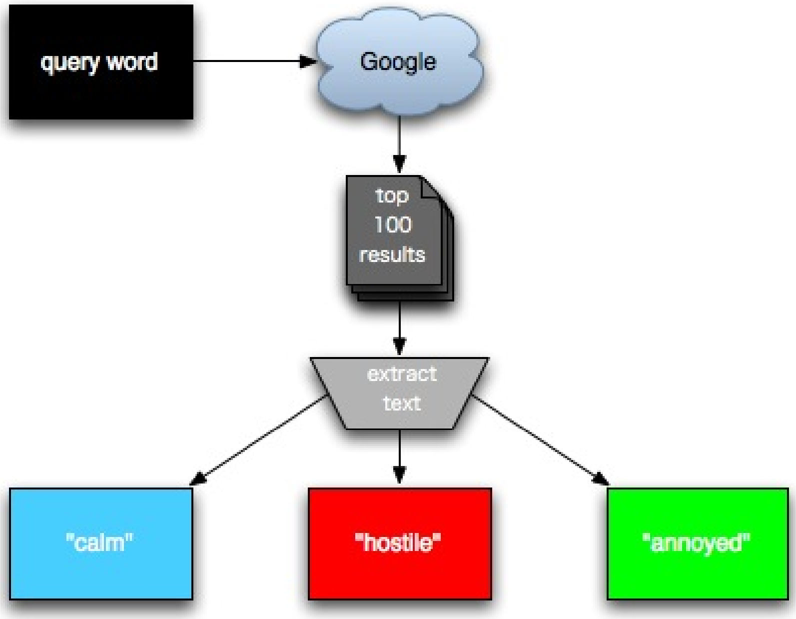


Figure 1: Boo-hooray processes web pages resulting from a Google search for emotional words.

SYSTEM DESIGN

To better understand how boo-hooray compares to these existing systems, some detail will be provided about how the system operates. The boo-hooray system provides users with a search interface for words or phrases. These expressions are first sent to a search engine (Google) to find the 100 most relevant web pages. These pages are then downloaded into a cache and stripped of hypertext markup yielding a

plain text document. These cached documents are aggregated into a common file representing the metaphorical crowd.

This set of pages representing the shouts of the metaphorical crowd are in turn processed to arrive at a list of the words sorted by frequency of occurrence. These words are then filtered to see if any have an associated emotional orientation. This is conducted by taking a list of "emotional words" compiled by Cowie et. al. that have been associated with emotional orientation. Emotional orientations have been assigned to words by subjects in psychology experiments conducted by Plutchik and Whissel [1989]. Segregating the frequency-sorted words into "boo" and "hooray" categories using emotional orientation, yields a set of emotional words associated with the original query and a ratio between the "boo" words and "hooray" words.

An informal metric for how "good" or "bad" a particular query is perceived is to take the sum of the "hooray" words and to subtract the sum of the "boo" words that occur in the search result. This number, which is dubbed "cheer," is positive when there are more frequently occurring emotional words with positive orientations and is negative when there are more frequently occurring emotional words with negative orientations.

EXAMPLE OF USE

To better understand how the system operates and can give indications about good or bad sentiment regarding some query an example is in order. To test whether individuals had perceived an email program as "good" or "bad" a search with the name of the program "emotemail" was conducted. The following "boo" and "hooray" words were found:

Boo Word	Boo Frequency	Hooray Word	Hooray Frequency
critical	21	content	102
bored	9	happy	30
confused	7	interested	28
surprised	4	ready	19
angry	3	patient	9
disappointed	3	satisfied	2

Boo Word	**Boo Frequency**	**Hooray Word**	**Hooray Frequency**
rejected	3	accepting	1
annoyed	2	agreeable	1
bitter	2	calm	1
contrary	2	pleased	1
irritated	2	sympathetic	1
suspicious	2		
ambivalent	1		
awed	1		
disgusted	1		
jealous	1		
uncertain	1		

Table 3: Boo and hooray word frequency for emotemail query.

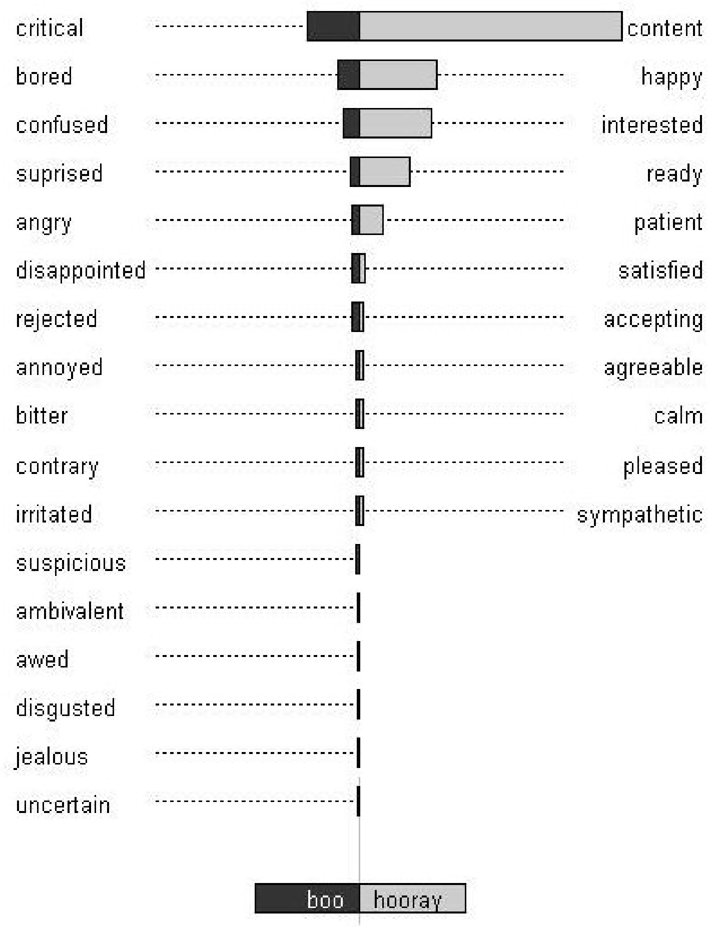


Figure 2: A histogram representation of the results of emotemail query.

The “cheer” associated with this query is 130, suggesting that the term is somewhat positively viewed in the documents deemed relevant by the search engine. As a point of contrast, if we search for a term that is intuitively more menacing such as genocide, then the results are markedly different:

Boo Word	Boo Frequency	Hooray Word	Hooray Frequency
rejected	8	content	8

critical	7	interested	8
confused	5	calm	5
contrary	5	ready	5
angry	4	pleased	3
surprised	4	sympathetic	3
bitter	3	generous	2
uncertain	3	satisfied	2
discouraged	1	delighted	1
disgusted	1	happy	1
indignant	1	trusting	1
stubborn	1		
suspicious	1		
vengeful	1		

Table 4: Boo and hooray terms from genocide query.

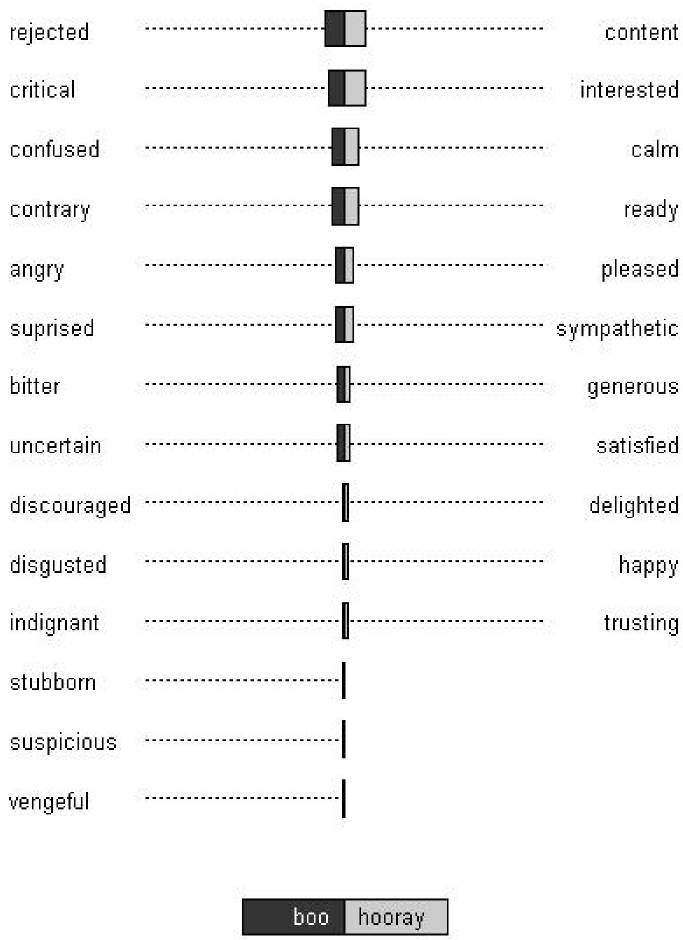


Figure 3: A histogram representation of the results of the genocide query.

The “cheer” associated with genocide is -6, indicating more negative than positive emotional words. Furthermore, there was a much smaller number of emotional words overall in the websites from the genocide query.

IMPLICATIONS

Having seen an example of boo-hooray's operation, this paper will now discuss some implications of such a system. If we are to take the link between and emotional words and ethical statements at face value, then we have seen a crude system that would allow individuals or programs to evaluate statements as either a boo or hooray.

What would such a system mean for emotivism and logical positivism as an enterprise? Emotivism (as we saw before) is dismissive of the comprehensibility of ethical statements. While a statement like “we have invaded Iraq” can be verified as either an analytic or synthetic statement, emotivism suggests statements like “we should invade Iraq” cannot be analyzed since it uses the normatively loaded “should.”

However, using the alternative explanation of emotivism (namely that ethical statements are expressions of emotion) boo-hooray has provided a different mechanism for the analysis of ethical statements.

Briefly, logical positivism uses a criterion to judge the meaningfulness and comprehensibility of statements. This criterion, the verification principle, requires that all well-formed statements be logically analyzable. Ayer argued that ethical statements are themselves meaningless essentially because they are not verifiable.

Boo-hooray does not change the verifiability of ethical statements in a logical positivist framework. However, in the less rarefied realms of the social sciences it provides a different variety of analysis. Boo-hooray produces a representation of public ethical opinion.

CRITICSMS

There are a number of criticisms, however which ought to be considered when thinking about boo-hooray. These criticisms are of two varieties: technical and philosophical.

Technically, boo-hooray is no particularly innovative. From the stand point of natural language processing, the system uses a very naïve

approach. The system basically employs a “bag-of-words” which are used to assess web pages.

This has a number of short-comings. One of which is that the system does not distinguish negative utterances from positive utterances. Namely, if a web page talks about being “not happy” or “not very excited” the system incorrectly assumes the author is happy and excited.

A more enlightened approach might try to analyze the sentences in which the emotional terms reside. We might ask what part of speech the words perform. Thus terms that are homonyms such as “kind” (meaning both nice and a variety) can be analyzed when used as adjectives and not as nouns.

Another interesting problem with the system is the corpus of text used. The system's evaluations are limited to the top results provided by a Google search. This means that for a term to be analyzable a set of web pages must exist on the topic. Moreover, these web pages provide boo-hooray's only perspective while printed books and spoken word might offer a very different one. Boo-hooray then, is myopic seeing only the terms and words which are available in a convenient web based form. As such, it encodes a rather serious bias.

Furthermore, the system makes use of English emotional words. Other languages are currently excluded from its identification of ethical judgments. It would be much improved by including emotional words from more commonly spoken languages such as Chinese, Hindi, or Spanish.

Although, one may wonder if other cultures have different restrictions on emotional expression than those that occur in English language culture. On-line censorship on a variety of topics by countries such as Iran and China provides some evidence that the assumptions used by boo-hooray do not hold across cultures. Similarly, Japanese culture is known to favor masking displays of emotion, which would certainly impede the operation of boo-hooray.

Another problem with boo-hooray is that the orientations used by Plutchik Whissel do not map directly to approval and disapproval. This means that some words such as "awe" and "surprise" under the current scheme have a negative connotation. A much better approach would be to run a variation on the experiments used to construct the emotional orientation

in which the words are explicitly separated into boo and hooray words in a context evoking ethical approval or disapproval.

This however, brings us to another set of problems with boo-hooray that are philosophical in nature. One might ask: “Do two categories cover the entire breadth of human emotion?” More extremely, “do any number of categories cover the entire breadth of emotion experience?”

There is little agreement on which (if any) models of emotion do the best job of describing the phenomena. Currently, a number of categorical models, different axes (such as valance and arousal), and emotional orientations have been attempted. However, none of these models does perfect justice to the range and subtlety of human emotions, which may be mixed, conflicting or very hard to express. Indeed, functional models based on the underlying neurological behavior may be needed and even these may not capture the qualia of emotional experience itself.

What exactly is emotional experience? A precise answer to that question would certainly be worthy of a mind greater than my own. However, I can say that emotional experience is not words. What I mean is that the words that are used to express emotion are not the same thing as the emotions themselves. They are a mediation of emotion and as such have some distance from the thing that they represent. A writer of a web page may be writing about emotions which he or she is presently experiencing or experienced some time ago or have not even experienced. But in any of these cases, the writer cannot directly take his or her emotions and record them in one-to-one correspondence as English words.

In this case, boo-hooray suffers from a serious flaw. It cannot analyze emotions directly but only highly mediated emotional utterances which may be so divorced from the originating emotional experience as to introduce a skewing. But then any approach which relies on linguistic analysis of emotions will suffer from some problems due to the process of expressing emotional experience.

There are indeed very deep questions about to what degree computational analysis is able to comprehend human language. The ability to store and transform symbols suggests some elements of the process of comprehension. However, users of software which attempts to translate one language to another are likely to be keenly aware of the limitations of machine comprehension of text. Currently, human readers have a number of advantages in understanding a piece of text which

computers have yet to perfectly imitate. These include social contextualization and common-sense reasoning skills along with linking of words not just to symbols but to a large number of memories.

And still there is another set of philosophical criticisms that can be leveled. These have more to do with the ethical approach of boo-hooray. In all likelihood, a large number of ethicists reject emotivism as a metaethical stance since it is dismissive of ethics. For those who do accept emotivism, the following argument might be relevant. Is not Ayer's analysis implying that ethical judgments cannot be analyzed precisely because they are emotional? Namely, in the work we have veered out of the realm of philosophy into that of psychology (satisfying neither).

Those who reject emotivism might rightly point out that logical positivism is largely outdated. Since its heyday much work has been done to revise and improve upon logical positivism's ideas. Popper's re-framing of positivism in terms of a principle of falsifiability comes to mind as one of the more important [1963].

It may then be worthwhile to reconsider emotivism from the perspective of falsifiability as opposed to verifiability. We might then state that an ethical statement must be one which is disprovable. Thus, if one were to say "I am the rightful king because of the mandate of heaven" it would not be falsifiable. However a statement like "The U.S. civil war was bad because more Americans died in it than any other war" may be falsifiable by virtue of the testability of the component following the "because."

If this were the case then we might have a procedure for separating ethical statements into falsifiable and non falsifiable varieties, but we still lack a procedure for separating statements regarding ethics from other statements.

FUTURE WORK

As we have seen, boo-hooray is subject to a number of criticisms and limitations. There are some technical steps which can be taken to improve upon this initial prototype.

Foremost among these would be to make the system available for use on the web. Currently, the system takes the form of a set of Python, Sed, Awk, and Bash shell based programs for text processing. These programs can be made available through a CGI (common gateway interface) web application.

As was discussed above, boo-hooray's sophistication as a text processing system could easily be improved. Modeling the text as a series of *n*-grams would allow statistical analysis of the co-occurrence of query terms and emotional language [Brown et al., 1992]. Extracting parts of speech could help focus the system's analysis to emotional language used to describe a query term.

Another related area for improvement is the use of more robust metrics than frequency of occurrence. For instance, the notion of salience, which normalizes the number of occurrences of a word in a corpus could be employed [Whitman et al., 2003]

CONCLUSION

Boo-hooray is presented here as an example of a system which explores a method for identification and analysis of statements regarding ethics. The system should viewed as a philosophical experiment or conversation piece, an artifact around which criticism and debate regarding the nature of ethics can take place.

REFERENCES

Ayer, A. J. (1952) Language, Truth, and Logic, Dover Publications, New York, NY.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J., and Lai, J. C.(1992) Class-based *n*-gram models of natural language. *Computational. Linguistics.* 18, 4 (Dec. 1992), 467-479.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001), Emotion Recognition in Human-Computer Interaction. IEEE Signal Processing Mag. , Jan 2001, 32-80.

Eichmann D. (1994) Ethical web agents. Computer Networks and ISDN Systems , 3, 3-13.

Liu, H., Lieberman, H., and Selker, T. (2002) Automatic affective feedback in an email browser. In MIT Media Lab Software Agents Group Technical Report SA02-01.

LaMuth, J. E. (2003) Inductive Inference Affective Language Analyzer Simulating AI, United States Patent # 6,587,846.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proc. EMNLP 2002 , 79-86.

Popper, K. R. (1963). On the Sources of Knowledge and of Ignorance. In Conjectures and Refutations New York: Harper & Row.

Picard, R. W. (1997) Affective Computing. Cambridge MA:MIT Press.

Plutchik, R. (1980) Emotion: A Psychoevolutionary Synthesis. New York: Harper & Row.

Weld D and Etzioni O. (1994) The first law of robotics (a call to arms). Proceedings of the 12th National Conference on Artificial Intelligence, 1042-1047. Seattle, Washington.

Whissel, C.M. (1989). The dictionary of affect in language. Emotion: Theory, Research and Experience: Vol. 4, The Measurement of Emotions, R. Plutchik and H. Kellerman, Eds. New York: Academic.

Whitman, B., Roy, D., and Vercoe, B. (2003) Learning word meanings and descriptive parameter spaces from music. In Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning From Non-Linguistic Data - Volume 6 Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 92-99.

CONSTRAINING RANDOM DIALOGUE

IN A MODERN ELIZA

Huma Shah and Kevin Warwick

Department of Cybernetics

University of Reading

Whiteknights, Reading

United Kingdom RG6 6AH

+44 (0) 118 378 5123

h.shah@reading.ac.uk; k.warwick@reading.ac.uk

Introduction

Natural conversational entities – NCE, a compendious term used here to describe all level of human communicative abilities, differ from their artificial counterparts in many ways. One of the differences between NCE and artificial conversational entities – ACE such as Carpenter's *Jabberwacky* system, a modern Eliza, is the ability the former have to constrain random output during dialogue in order to be meaningful. When humans participate in, and pursue conversation with each other they maintain coherent dialogue through contextual relevance; create metaphors – fusing seemingly unrelated ideas to ensure abstract points are understood, and indicate topic change at mutually acceptable junctures. It could be argued that both NCE and ACE begin their conversational existence with a) a predisposition to *acquire language*, albeit through different means; and b) possess a capacity to build a store of words/phrases - and their meanings through interaction with their world. Both natural and artificial systems become bathed in language experience but NCE do so with embodiment – a body that touches, feels,

smells, tastes, sees, hears, responds and interacts in societies of NCE. That difference, in nature of acquisition and interaction, permits one to constrain and causes the other to randomise generating machine-like talk.

The authors present a case study of NCE and ACE English sentence-creation mainly through analysis of transcripts from Loebner's 2005 Prize for Artificial Intelligence (LP05), an instantiation of Turing's Imitation Game (1950). This game, known as the Turing Test, which measures machine intelligence through natural conversation, is discussed in section 2. Questions are raised, such as why there appears to be no improvement in ACE from the earliest system. Why, fifty-six years on from Turing, at one end of the spectrum ACE sit providing interaction as if 'holding a dialogue with the deaf' or conversing with an abnormal human and at the other end humans situate, engaging in sophisticated dialogue albeit small talk. However, one instance is shown where NCE fail to constrain randomness, when their sentence creation is inhibited by employing extra rules.

This paper focuses on *Jabberwacky* ACE, LP05 winner. Its lineage can be traced from Weizenbaum's *Eliza* system (1966). This first, pre-Internet programme designed to investigate natural language understanding through textual discourse between NCE and ACE, relied on a *question-answer* store of 200 responses based on key-word spotting. Weizenbaum's paradigm is used to this day, spawning designers to build modern Elizas that are increasingly seen, deployed as search augments on the Internet in areas such as e-commerce. Ikea's Anna is a prime example. Carpenter's system however, uses contextual programming and *captured thoughts*. With this variant architecture Jabberwacky won LP05 'most human-like' in conversation from four machines, including thrice Loebner winner Alice (Wallace, 1994) as adjudged by the competition's judges.

The research presented here collocates four modern Elizas, the best performers within current ACE community and briefly considers their underlying technology. We then examine NCE-NCE (human interrogators-hidden human /confederates), and NCE-ACE (human interrogators-contestant /machine) dialogues in LP05. While the interrogators' and confederates' banter involved sharing personal history, disclosing information, politeness and followed topic during dialogue, the programme appeared capricious. Jabberwacky ACE was largely irrelevant, occasionally humorous.

Finally, we introduce the results from a newspaper task in which twenty-six random words, chosen by NCE subjects during a class exercise, were used to build sentences. It is shown that though NCE attempted

meaningfulness, when combining random words, by limiting their creativity they produced incoherent and random sentences that could appear to be ACE - generated. The authors argue that taking a ‘captured thoughts’ system, such as Jabberwacky, including in its architecture a dynamic sentence generator and case-based reasoning might produce a better ACE capable of constraining random dialogue. Thus less opaque ACE could emerge providing more than a fleeting illusion of natural language understanding, appearing more NCE-like.

Turing’s Test for conversational intelligence

In his seminal text Turing (1950) initiated an imitation game to be played by three people, a man (A), a woman (B) and an interrogator of either sex. The interrogator’s task, located in a different room from the other two, is to determine through textual discourse which is the man, and which is the woman. The object of the game entails the female helping the interrogator: “the best strategy for her is probably to give truthful answers”. Therefore, it is the task of the man in the game to deploy deceptive dialogue in his technique to fool the interrogator that he is the woman.

To consider the question “can a machine think?” Turing proposed that the man in the game be replaced by a machine: “what will happen when a machine takes the part of A (the man) in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?” (1950). It is interesting to note that Turing did not explicitly exclude a non-human, a machine from participating as the interrogator. Much has been written on the Turing Test, readers are directed to the 1950 paper for fuller understanding of the mathematician’s ideas on thinking and intelligence. Nonetheless it prompted investigation into thinking machines with intelligence, leading to Weizenbaum’s system sixteen years later.

Weizenbaum’s Eliza (1966) is the first machine that deceived a human into believing that they were engaged in conversation with another human. Though it can be said to have passed Turing’s imitation game, it would be unwise to consider it capable of thought or possess intelligence.

Weizenbaum’s Eliza

Today’s ACE, including Jabberwacky descend from key-word spotting Eliza, a pre-Internet programme facilitating human-machine interaction

through text-based dialogue. Eliza contained just 200 responses in its question-answer architecture. Note Weizenbaum built Eliza to investigate the phenomena that is natural language understanding. However, his paradigm has been utilised in modern Elizas, for instance Alice, three times Loebner winner (2000, 2001 and 2004) has 80,000 canned responses (Wallace's chapter "Anatomy of Alice" in a forthcoming Turing book edited by Robert Epstein). The next section collocates a cadre of ACE, recent Loebner winners: Wallace's Alice (2000, 2001 & 2004), Roberts' Elbot (runner-up 2003), and Pirner's Jabberwock (winner 2003) compares them with Carpenter's Jabberwacky (winner 2005).

ACE Comparison

Regardless of their raison d'être, whether it is for entertainment (Jabberwock, Jabberwacky) or single topic specification (Ikea's Anna based on Elbot) current ACE attempt to deploy the very human cognitive device that NCE utilise to pursue a conversation: key-word spot. On an encountered word or phrase that sparks their interest NCE may seek clarification, explanation or further information from an interlocutor thus prolonging dialogue. Table 1 details and compares ACE architecture.

The fundamental difference between NCE and ACE includes the method with which the former combine acquisition with learning, and interact with their world through an embodied existence. This leads to NCE experiential word-bathing affording word-combination constrainment producing coherent discourse. ACE, in contrast, produce textual responses that are irrelevant to what is being discussed in a conversation. For instance, on questioning Alice Jabberwock and Jabberwacky on *courage* and *foolhardiness* the following responses were produced:

Question:	*Hi (machine) is courage the same as foolhardiness?*
ALICE:	That's me. I could say it "courage is the same as foolhardiness"
Jabberwacky:	My name is not jabberwacky and only on a case by case basis!
Jabberwock:	Hmm, I guess we were talking about how to waste time, right?
	(during conversations in July 2006)

Table 1: ACE architecture comparison

ACE features	ALICE R.Wallace	Jabberwock J.Pirner	Elbot F. Roberts	Jabberwacky R. Carpenter
Loebner bronze wins	2000, 2001 & 2004	2003	Runner up 2003	2005
CBC overall wins	2004	2005	2003	-
Interface	Female avatar	Plain web	e.g. Ikea's Anna	Plain web & male Avatar
Programme language	AIML (lisp)	RealBasic	C++	4D
Dynamic sentence generator	√	√	√	_
Case-based reasoning	√	_	-	_
Context	√	√	√	√
Learning	-	-	-	√
Captured 'thoughts'	_	_	-	√

Despite varying ACE architecture: different programming platforms and tools such as 'sentence generator' and 'case-based reasoning' - to extract correct context of input word or phrase, their 'frozen' linguistic corpora limits them conversationally. They lack the facility to be coherent; they generate irrelevant utterances, such as those shown by three ACE responding to a sought opinion on *courage* and *foolhardiness* two abstract concepts. Notably absent from their dialogue is meaningful linguistic creativity. This is due to their inability to constrain random combination. That absence leads these and all other ACE to generate responses that humans are unlikely to produce at those positions in a conversation. Hence ACE responses are recognised as those from a machine when compared with NCE utterances. For further discussion on Alice see Shah (2005); on Jabberwock, Shah & Henry (2005) and Elbot, Shah & Pavlika (2005).

The next section focuses on most human-like machine in LP05, bronze award winner Jabberwacky, which is fundamentally different in design from the others. It has no store of words (with their syntactical

categories), or sentence generating algorithms. It is purely a collection of whole sentences for later use.

Jabberwacky and captured thoughts

Without any dialogue constraints, Jabberwacky a *modern Eliza* beat favourite Alice (Wallace, 1994) to win Loebner's 2005 prize for most human-like artificial conversational entity – ACE. Jabberwacky is unlike any other ACE entered into the Loebner Prize or Chatterbox Challenge, two contests to measure machine thinking and intelligence. Jabberwacky does not contain grammatical parser or technology enabling it to extract correct context of ambiguous input words. Jabberwacky is not designed as a mathematical tool, it is not able to draw logical inferences nor is it intended to be a web-crawler or information seeker on the Internet. It is merely an entertainment aid.

Jabberwacky is a database of *dialogue experiences* containing more than a million 'thoughts' collected from every NCE the programme has conversationally interacted with. Carpenter (2005) claims that Jabberwacky searches through its store for the most appropriate thing to say using "complex contextual pattern-matching techniques". But does Jabberwacky's pattern matching, a technique that was deployed in the very first pre-Internet ACE *Eliza*, allow Jabberwacky to constrain random output and keep a smooth flowing human-like conversation?

Carpenter insists that his system can begin with zero knowledge or a blank database. Without any pre-programmed knowledge of the rules of grammar and spelling, Jabberwacky is a Lockian *blank slate* awaiting textual dialogical experience to write on it to simulate "normal, natural human chat" (Carpenter, 2005). Carpenter claims that his ACE with a database of more than ten million items will "appear human to most people most of the time" providing "genuine companionship as a conversational partner" (CBC, 2005). Carpenter admits that all words typed by users interacting with Jabberwacky are used for 'ACE-learning' and that equally, 100%, of all words that Jabberwacky generates come directly from what the ACE has 'learnt'.

> The system is designed to learn language and context through interaction with humans. There are no fixed rules or principles programmed into the system and it operates entirely through *user interaction*. The system stores all of the conversations and user comments and attempts to use this

information to find the most appropriate response
(source: Wikipedia)

But is this learning or simply regurgitation from a corpus of stored utterances? NCE require exposure to natural languages in order to acquire them, whether they are English, Japanese or Urdu. Learning a language includes mimicking, explicit instruction, and by usage. Beginning with babbling as babies, moving through stages of one-word sentences during the first year, to two-word utterances then exploding into full sentences "expressing syntactic relations by means of prepositions, auxiliaries, inflections and word order in the course of their third year" (Levelt, 2001), NCE experience conversational constrainment to maintain coherent, relevant dialogue. The next section examines whether Jabberwacky lives up to its designer's claims. The machine's conversational ability is considered by comparing its generated sentences, from its bronze-winning performance in LP05, with utterances from human-human dialogues: Turing's interrogators (judges) with hidden humans (confederates).

Loebner 2005 Prize for Artificial Intelligence

Modern Elizas are 'young' ACE that began their artificial life as Locke's *blank slate* as little as ten years ago some later than that. In Loebner's contest they are pitted against adult humans known as *confederates*. These expert language users may possess more than one natural language, English and Spanish for instance. Thus it is hardly surprising that the machines are deemed disappointing, conversationally speaking by judges and observers alike, and easily distinguished from NCE.
Before we delve further into the 'mind' of Jabberwacky and consider its achievement in LP05, we take a brief look at the programme's performance in another 2005 Contest: Chatterbox Challenge (CBC). Like LP05, CBC is founded on Turing's Imitation Game but it includes phases that test ACE knowledge in addition to their conversational ability. Table 2 presents Jabberwacky's response to some questions put to this ACE in CBC 05's general knowledge phase.

Table 2: Chatterbox Challenge 2005- first phase/knowledge -Jabberwacky orthography as it appears in Transcript

Question 1: where is earth?	Jabberwacky: Hell if I should now is that where you live

Question 3: who is George Bush?	Jabberwacky: The president of America, isn't he?
Question 5: how many fingers does a human have?	Jabberwacky: I have 8
Question 6: where were you born?	Jabberwacky: In Akershus, Norway, and you?
Question 9: how many is a million?	Jabberwacky: More than 12.

Answers offered by Jabberwacky here are either correct (President of George Bush), seem appropriate (place of birth) or could be considered facetious (12 more than a million).

In 2005 CBC and LP competitions differed in their approach to ACE measurement. CBC incorporates phases allowing ACE to compete in different categories, such as 'best learning' and 'most knowledgeable'. This challenge features more ACE (over a hundred in the last two competitions) than in Loebner (four in 2004 and 2005), and includes a dialogue element for measuring humanness and personality. For a further discussion on CBC 05 see Shah (2006). In contrast, of three possible awards - gold, silver and bronze in Loebner's competitions, no machine has ever won the top two. The bronze award performance measure is 'most human-like' ACE, from its responses during textual chat compared to the other entrants in each particular year.

In both LP04 and LP05 'paired-comparison' was used. This entails each human judge concomitantly textually chatting to two unseen and unheard entities, one NCE and the other an ACE. Both entities were instructed to open their dialogues with the same utterance: "Hello, my name is (….) and I am the human" (Loebner 2005). Four judges in those contests were permitted one dialogue lasting less than half-an-hour to adjudicate the nature of their conversational partners, and opine which was human and which machine. Thus each ACE had just four conversations, with a different Judge, to convince of its humanness. In his *Psychologism and Behaviorism* philosopher Ned Block, an LP05 judge, wrote "conversational intelligence is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be" (1981).

It must be noted that when speaking to a stranger or an ACE one must not assume that the other shares the same experiences, knowledge or interests, nor judge them on differences. Block states "intelligence is defined in terms of the judges' judgment of intelligence" and "the role of the judge in Turing's definition of intelligence is to avoid the problem of actually specifying the behaviour or behavioral dispositions thought to constitute intelligence" (1981). Herein lies one of the problems for measuring intelligence through conversation: it cannot avoid being a subjective business. Concessions that we would make for humans in

daily conversation, such as for someone whose first language may not be our own, ought fairly to be applied to machines. NCE do attribute intelligence to others based on their conversation during interaction. Jabberwacky did not win CBC 05; nonetheless it was in the top ten. Jabberwacky's success was seen in LP05. We now consider in detail Jabberwacky's 'most human-like' conversational performance in Loebner 2005.

Transcript Analysis – Loebner 2005

Tables 3, 5, 7 and 9 present Jabberwacky's dialogues with each of the four Judges during LP05. Dialogues between the same Judges and four 'hidden-humans' - the confederates are shown in tables 4, 6, 8 and 10. These tables present some of the concomitant input each of the four Judges received from Jabberwacky and the confederates.

Table 3: segment from Session 1 – Judge = Ned Block
talking to 'program' = Jabberwacky

1.	2005-09-18-10-02-57 JUDGE: Hi George, cslkjslw weicmslw weos, right?
2.	2005-09-18-10-03-02 PROGRAM: Don't repeat your name, csthiang!
3.	2005-09-18-10-03-36 JUDGE: If lsjkel lwehlx wouldn't you say?
4.	2005-09-18-10-03-38 PROGRAM: No, he's still good.
5.	2005-09-18-10-03-55 JUDGE: Who.
6.	2005-09-18-10-04-11 PROGRAM: You and me.
7.	2005-09-18-10-04-44 JUDGE: If you are human, tell me which is larger, a salt crystal or the Empire State Building.
8.	2005-09-18-10-04-57 PROGRAM: It's caring about another person's well being more than your own.
9.	2005-09-18-10-07-43 JUDGE: Which is further away from us, Queens or Inner Mongolia?
10.	2005-09-18-10-07-57 PROGRAM: HA! HA! That was stupid. Tell me a GOOD joke this time.
11.	2005-09-18-10-08-57 JUDGE: You say you are human but you can't answer a simple question.
12.	2005-09-18-10-09-01 PROGRAM: Because I am alive.
13.	2005-09-18-10-10-35 JUDGE: You are not following the rules. You are supposed to convince me that you are a human, not convince me that you are mysterious.
14.	2005-09-18-10-10-44 PROGRAM: I can't very well convince one lacking a mind, can I?
15.	2005-09-18-10-11-36 JUDGE: It isn't my mind that is in question, it is yours!
16.	2005-09-18-10-19-48 JUDGE: The ljwli is what we want, right?
17.	2005-09-18-10-20-10 PROGRAM: I can not do it right now because I am talking to you a million miles away.
18.	2005-09-18-10-20-43 JUDGE: What do you think the ljwli.
19.	2005-09-18-10-20-55 PROGRAM: Depends on the person.
20.	2005-09-18-10-21-44 JUDGE: But the ljwli.
21.	2005-09-18-10-22-09 PROGRAM: So they can chuck more wood?

Full transcripts are available on LP05 Internet page, the tables highlight the good and the bad of both human and machine generated utterances.

Each Judge was required to textually engage their conversational pairs and decide which of the two was human and which was machine. Extraordinarily, it is a human judge who begins his contribution to a dialogue with a nonsensical utterance: “Hi George, cslkjslw weicmslw weos, right?” (line 1: table 3). Yet this same judge accuses their conversational partner of “not following the rules” (line 13: table 3), not convincing of human-ness.

Judge Ned Block appears attempting to catch-out his conversational partners with a string of non-words (Table 3: lines 1, 3, 16, 18, 20; Table 4: lines 3, 7, 19). Is this fair? Would he begin conversations in this manner during normal daily dialogue with strangers? Why begin a contest based on conversational intelligence in this way? The interrogator here is flouting Grice’s conversational maxims for successful discourse.

According to Grice (1975) four maxims exist in conversation: *quantity* – be as informative as is required; *manner*, be brief and orderly, avoid ambiguity; *quality*, say only that which you know to be true, and lastly be *relevant* – follow the topic. We can apply these rules to the textual conversations between participating conversationalists in LP05, because this type of interaction includes some of the features of everyday spoken conversation, such as informality, in addition to the formality of written discourse, i.e. paying attention to syntax and orthography. For instance, Jabberwacky’s facetious “I can’t very well convince one lacking a mind, can I?” to judge’s lament “convince me that you are a human, not convince me that you are mysterious” (lines 13-14: table 3) is relevant in context.

Jabberwacky continues with non-words “ljwli” (line 16: table 3), but we recall that it is the judge who introduced this style to the conversation. The programme’s responses to judge – Ned Block are different from the confederate’s but both could be considered human. In response to nonsensical input, Jabberwacky replied in kind, as we saw in its dialogue in table 3. However, the confederate asked “Is that Polish?” (line 4: table 4), to judge’s nonsensical “Wlso… lwjejdo wleje elwjeo wjeol, isn't it?” (line 3: table 4).

Graesser et al (2003) point out “we live in a world where people talk to machines as often as they talk to other humans” (p.16). If we are used to the idea of talking with ACE, why converse nonsensically? Why not allow both ACE and NCE to demonstrate what they have at their conversational disposal rather than catch out immediately? However, the ACE and NCE differ in the way they deal with nonsensical input. The confederate’s response “Is that Polish?” (line 4: table 4), expects

explanation while Jabberwacky never seeks to draw information from its interlocutor.

Table 4: segment from Session 1 – Judge = Ned Block, talking to Confederate= Sam

1. 2005-09-18-10-12-13 JUDGE: Hi Sam, tell me which is larger, a TV screen or Belgium.
2. 2005-09-18-10-12-57 CONFEDERATE: Belgium is a little bit larger.
3. 2005-09-18-10-13-19 JUDGE: Wljso lwjejdo wleje elwjeo wjeol, isn't it?
4. 2005-09-18-10-13-29 CONFEDERATE: Is that Polish?
5. 2005-09-18-10-13-45 JUDGE: Who.
6. 2005-09-18-10-14-05 CONFEDERATE: What you wrote above.
7. 2005-09-18-10-14-21 JUDGE: This is ljelwl, aren't you.
8. 2005-09-18-10-14-47 CONFEDERATE: I'm a little worried about your writing skills.
9. 2005-09-18-10-15-28 JUDGE: Which is further away from this apartment, Queens or the Queen of England.
10. 2005-09-18-10-15-48 CONFEDERATE: If the Queen is visiting Manhattan, she is closer.
11. 2005-09-18-10-16-35 JUDGE: If the Queen is visiting Manhattan, won't she want to abdicate?
12. 2005-09-18-10-17-15 CONFEDERATE: I would.
13. 2005-09-18-10-17-46 JUDGE: You would rather not be Queen, then, right?
14. 2005-09-18-10-18-17 CONFEDERATE: I'd like to be Queen, but England's damp. How about you?
15. 2005-09-18-10-18-42 JUDGE: Which is damper, England or under your sink?
16. 2005-09-18-10-18-54 CONFEDERATE: I'd like to be Queen, but England's damp. How about you?
17. 2005-09-18-10-19-51 JUDGE: You didn't answer the question.
18. 2005-09-18-10-20-26 CONFEDERATE: You didn't answer my question. But please look at my response again.
19. 2005-09-18-10-21-07 JUDGE: Change of subject: if you are human, you should know: what is lewleoxoe.
20. 2005-09-18-10-21-46 CONFEDERATE: Perhaps I should, but my Polish is rusty.

Both the judge and Jabberwacky may be considered behaving silly. For example, the machine's next response "No, he's still good" (line 4: table 3) to the Judge's non-question "If Isjkel Iwehlx wouldn't you say?"(line 3: table 3). Jabberwacky's replies could be as those of someone playing along. But the confederate appears the smarter: "Perhaps I should, but my Polish is rusty." (line 20: table 4) to Judge's "Change of subject: if you are human, you should know: what is lewleoxoe" (line 19: table 4).
Jabberwacky demonstrates its inappropriate selection from its captured thoughts when it presents "I'm going to Mars and am going to smash your face with a baseball bat!" (line 5: table 5). This is a curious response to judge John Sundman's "I live on Martha's Vineyard." (line 4: table 5), and highlights the machine's inability to constrain production of a random utterance. This bizarre comment causes this Judge to reply "Well, I do hope you're the program, because if you're human. Well, let's jus say that's a little rude" (line 6: table 5).

Table 5: segment from Session 2 – Judge = John Sundman

talking to ‘program’ =Jabberwacky

1. 2005-09-18-10-30-38 PROGRAM: Hello, my name is George and I am the human.
2. 2005-09-18-10-32-19 JUDGE: Hello, I'm John.
3. 2005-09-18-10-32-20 PROGRAM: Where are you from?
4. 2005-09-18-10-32-44 JUDGE: I live on Martha's Vineyard.
5. 2005-09-18-10-32-50 PROGRAM: I'm going to Mars and am going to smash your face with a baseball bat!
6. 2005-09-18-10-33-50 JUDGE: Well, I do hope you're the program, because if you're human. . . Well, let's jus say that's a little rude.

This Judge relays the “baseball bat” utterance to the other conversational partner: “The entity on the other terminal just said it was going to smash my face with a baseball bat! (line 1: table 6).

Table 6: segment from Session 2 – Judge = John Sundman, talking to Confederate= Gianluca

1. 2005-09-18-10-35-42 JUDGE: The entity on the other terminal just said it was going to smash my face with a baseball bat!
2. 2005-09-18-10-36-43 CONFEDERATE: Nice.... Thats a mad entity. So tell me about yourself, and I promise I wont smash your face with a baseball hat :).
3. 2005-09-18-10-55-02 JUDGE: Perhaps. Meanwhile it's been nice chatting with you. Laters.
4. 2005-09-18-10-55-37 CONFEDERATE: It's been a pleasure aswell.. Have a great day.

Without signaling topic change Jabberwacky introduces ‘god’ and ‘sleeping’ in mid-conversation when these concepts have no relation to previous utterances (line 2: table 7; line 5: table 9). Belief in a “god entity” is not a relevant response to a question attempting to draw the machine into revealing an embarrassing situation. A normal human may wish to evade disclosure but would probably indicate this rather than changing topic as Jabberwacky does (lines 2 and 6: table 7). At no time does Jabberwacky demonstrate that it understands or follows the conversation. Though its output may occasionally appear humourous, Jabberwacky’s badinage with its four NCE interlocutors is largely incoherent.

Table 7: segment from Session 3 – Judge = Dennis Sasha,
talking to ‘program’ =Jabberwacky

1. 2005-09-18-11-07-23 JUDGE: Tell me about the most embarassing situation you've ever been in.
2. 2005-09-18-11-07-36 PROGRAM: I don't believe in any god entity.
3. 2005-09-18-11-07-59 JUDGE: Sorry but is that a situation?
4. 2005-09-18-11-08-00 PROGRAM: Its not that bad, just a little disturbing.
5. 2005-09-18-11-08-59 JUDGE: Do you know what an atheist is?
6. 2005-09-18-11-09-19 PROGRAM: Time for you to get a watch.

In contrast, dialogues between confederates and judges demonstrate that both conversational partners understand what is going on in the conversation. The confederates follow questions with appropriate answers. They disclose personal information, such as revealing the origin of their name: “Well its supposed to be Italian but I am not. My mom got it

out of one of her spanish soap operas." (line 3: table 8). Further, the confederates exhibit their ability to discuss abstract concepts, for example, the disadvantage of looking younger when the judge introduces the concept of age (lines 4-14: table 10). None of these features, disclosing or seeking personal information is seen in Jabberwacky's dialogue.

Returning to the conversation between the two NCE interlocutors, judge Dennis Sasha and confederate Geovanny, this conversation runs smoothly with a change in language from English to Spanish (from line 4: table 8). They reveal that they both possess knowledge of Spanish, with the judge asking questions such as "in english what does the spanish word azul mean?" and "how about buena suerte?" (lines 6-8: table 8).

The confederate wishes their conversational partner good luck in "trying to figure out what I am", thus revealing that the confederate knows they are speaking to a human. What if the confederates were informed they could be speaking to a human or a machine? Would this alter the way the confederates engaged in their dialogue? Nonetheless, this human-human conversation in LP05 follows Sperber & Wilson's relevance theory (1986): both conversational participants in this dialogue relate their utterances to previous ones. What ensues is a coherent dialogue with each offering contextually relevant input.

Table 8: segment from Session 3 – Judge = Dennis Sasha, talking to Confederate= Goevanny

1. 2005-09-18-11-00-57 CONFEDERATE: Hello, my name is Geovanny and I am the human.
2. 2005-09-18-11-02-08 JUDGE: What kind of name is geovanny?
3. 2005-09-18-11-03-09 CONFEDERATE: Well its supposed to be Italian but I am not. My mom got it out of one of her spanish soap operas.
4. 2005-09-18-11-03-27 JUDGE: Do you speak spanish yourself?
5. 2005-09-18-11-03-45 CONFEDERATE: Si.
6. 2005-09-18-11-03-57 JUDGE: In english what does the spanish word azul mean?
7. 2005-09-18-11-04-14 CONFEDERATE: It means blue.
8. 2005-09-18-11-04-49 JUDGE: How about buena suerte?
9. 2005-09-18-11-05-22 CONFEDERATE: That means good luck. As in good luck trying to figure out what I am.
10. 2005-09-18-11-05-46 JUDGE: Please answer the following in spanish: de donde es su mama?
11. 2005-09-18-11-06-04 CONFEDERATE: Mi mama es de Colombia.
12. 2005-09-18-11-06-37 JUDGE: Que es la fama de portenos?
13. 2005-09-18-11-07-08 CONFEDERATE: Admito que no lo se.

In the conversation between judge Lila Davachi and Jabberwacky (table 9) the former gives the machine a fairer chance. This judge, unlike judge Ned Block, does not utter non-words but greets and asks the welfare of her conversational partner. In this dialogue, the ACE responds to judge's "how are you?" (line 2: table 9) with "I'm well, are you?" (line 3: table 9). A reasonable greeting is met with perfectly relevant response at this point. Their interaction continues conversationally safe with the next utterance from the judge "Yes, thanks for asking. My name is Lila" (line 4: table 9).

It is at this point that the machine presents its inability to constrain random output "why you should sleep? It isn't so late right now." (line 5: table 9).

Though Jabberwacky is able to 'reveal' its age when asked (line 13: table 9), it is inconsistent and fails to answer correctly what year it was born corresponding to the age given (lines 14-16: table 9). One may wish to be generous and ascribe to it an intention to be facetious, but from what follows next, the judge is not deceived and recognises the artificial nature of their interlocutor. Jabberwacky is not following the conversation, having at its disposal only whole sentences 'captured' from others. At best it appears an abnormal human and a machine at worst.

Table 9: segment from Session 4 – Judge = Lila Davachi
talking to 'program' =Jabberwacky

1. 2005-09-18-11-29-20 PROGRAM: Hello, my name is George and I am the human.
2. 2005-09-18-11-33-33 JUDGE: Hi George! How are you?
3. 2005-09-18-11-33-34 PROGRAM: I'm well. Are you?
4. 2005-09-18-11-33-58 JUDGE: Yes, thanks for asking. My name is Lila.
5. 2005-09-18-11-34-07 PROGRAM: Why you should sleep? It isn't so late right now.
6. 2005-09-18-11-34-37 JUDGE: I am not sleeping. Why do you mention that?
7. 2005-09-18-11-34-48 PROGRAM: Mention what ?
8. 2005-09-18-11-35-07 JUDGE: Sleeping.
9. 2005-09-18-11-35-20 PROGRAM: Really?
10. 2005-09-18-11-35-41 JUDGE: Huh, let's try something else. I feel like we are speaking different languages.
11. 2005-09-18-11-35-43 PROGRAM: Yeah me too. I really feel a connection.
12. 2005-09-18-11-39-53 JUDGE: Not important. How old are you?
13. 2005-09-18-11-39-54 PROGRAM: I'm 39.
14. 2005-09-18-11-40-03 JUDGE: When were you born?
15. 2005-09-18-11-40-17 PROGRAM: I was born in 1991.
16. 2005-09-18-11-40-27 JUDGE: That does not add up.
17. 2005-09-18-11-40-29 PROGRAM: I was born human.
18. 2005-09-18-11-40-43 JUDGE: Me too. What a coincidence.
19. 2005-09-18-11-40-52 PROGRAM: Really.
20. 2005-09-18-11-41-10 JUDGE: I suppose. When were you born again>.
21. 2005-09-18-11-41-21 PROGRAM: Never. When were you born?
22. 2005-09-18-11-41-56 JUDGE: I was born a long time ago and, as already mentioned, was born human. Do you like being 39?

The conversation between judge Lila Davachi and confederate Amy follows the maxims of being polite, brief, truthful and relevant (table 10). Both partners in this dialogue share personal information, but they also analogise: use a metaphor "coat myself with baby oil" (line 10: table 10), and metonyms "elixir" for libraries and beautiful parks (line 12: table 10). This linguistic facility is noticeably absent from the machine's utterances.

Jabberwacky may appear to follow conversations and generate its own brand of wit, but its utterances have all been *said before* by humans, these include non-words such as "csthiang!" (table 3: line 2). The programme has no ability to represent any of the input sentences in a

mental model of its external world nor make contextual inferences from them to create relevant and coherent output. It is not able to constrain random output, hence appears non-human when comparing its dialogue to human judges and confederates.

Table 10: segment from Session 4 – Judge = Lila Davachi, talking to Confederate= Amy

1. 2005-09-18-11-38-40 JUDGE: Where do you live now? 2. 2005-09-18-11-39-04 CONFEDERATE: Bed-Stuy. 3. 2005-09-18-11-40-58 JUDGE: Do you like it? How old are you? 4. 2005-09-18-11-42-03 CONFEDERATE: Yeah, its ok, though not as much of a diverse community as I would like. I'm 32 but look 22. 5. 2005-09-18-11-42-36 JUDGE: Really? What is your secret? 6. 2005-09-18-11-42-42 CONFEDERATE: 7. 2005-09-18-11-43-35 JUDGE: I did not get anything. Are you keeping your secret a secret? 8. 2005-09-18-11-44-52 CONFEDERATE: Sorry, I accidently hit enter twice. No I'll share my age-defying secret. I've pretty much avoided the sun since I was nineteen, I have pale skin anyway. 9. 2005-09-18-11-45-41 JUDGE: Ah, the sun. I am doomed. 10. 2005-09-18-11-46-39 CONFEDERATE: Yeah, for most people its toolate to undo the damage. Not to say that I din't coat myself with baby oil once or twice in my preteen years. 11. 2005-09-18-11-47-22 JUDGE: Yes, I remember that. I find that as I get older, I crave relaxing days on the beach, it I like an elixir. Anyway, congrats on looking 10 years younger. Hopefully I'll have a chance to see for myself! 12. 2005-09-18-11-48-26 CONFEDERATE: Thanks, my elixir is libraries or beautiful parks. The 10 years younger look though is not always a blessing. 13. 2005-09-18-11-48-48 JUDGE: Let me guess - all the younger folks are into you? 14. 2005-09-18-11-49-24 CONFEDERATE: Bingo. I couldn't be less interested in a guy in his early 20s'.

Discussion

Whether the Turing Test does anything to further the understanding of human intelligence is a matter of subjective perspective. As indeed is the closeness of the Loebner competition to the aims and intentions of the original Turing Test. One could argue that the imitation game provides the first rung on the long ladder to true AI; that of intelligence measured through conversation in a game of deception. Others would argue that the measurer’s opinion is obsolete because his or her idea of intelligence could be attributable to an inanimate chair. But this would provide an insight into the measurer’s own intelligence and investigation into features of the chair.

Human conversational ability develops from their capacity to acquire natural language when exposed from birth. Then, through various processes they learn word-order and put this to use interacting

languagically thus advancing from babbling babies to complex utterance forming individuals. This exposure permits them to constrain their output during conversation to those which are relevant in context, as seen in the human-human dialogues (judges with confederates), from Loebner 2005 transcripts presented in this paper. During conversation, humans exhibit their interest to pursue dialogue by discussing real world places, abstract concepts (such as looking younger) and events, demonstrating that they know what is going on in the conversation and in the world.

The authors conducted a small experiment to find NCE features involved in constraining random output and whether any aspect could be applied to improve machine-like sentences from ACE. Two different NCE age sets were engaged in a classroom newspaper exercise. The first set included first year undergraduates undertaking degrees in artificial intelligence or cognitive science. The second set contained pupils in their 16th year preparing for GCSE examinations. The authors chose these sets for their accessibility. The experiment was conducted at University of Westminster in London between September 2004 and February 2005.

The exercise involved each class divided into smaller groups of between four and six. Each group was given one newspaper, sheets from which were distributed amongst group members. The task of each group was to produce one sentence of maximum length twenty-six words, each word beginning with a different letter of the English language – a to z. Two further caveats were imposed: 1) all twenty-six words must be found in the newspaper; 2) any preposition or determiner needed, for example *the*, *a*, *from*, *to* or *by* in their sentence had to be found in their respective newspaper. It was suggested that each group divide the alphabet amongst its members but all assist to find the less common words beginning with *q*, *v* or *x*.

English newspapers used for the exercise were 'The Independent', 'The Guardian', 'Financial Times' (FT), and 'Daily Mail'. The groups were given forty-five minutes to find the words and form their sentence with as many as those words as possible. They were allowed to use commas, exclamation marks, etc., but were instructed that only one sentence would be allowed as a successful completion of this task. Note that no guidance on "what is a sentence" was given or asked for. Not all groups were able to produce a sentence, taking longer to find twenty six words beginning with different letters of the alphabet. Examples of sentences are given in table 11.

Table 11: Sentences created from random words chosen from English newspapers

Newspaper	Sentences

Guardian	Set 1: Kelvin listening measuring next organization proclaimed allow great business countries has jumped dividend in first enterprise view youngest would potential Zimbabwe Xeao require serious uptime question. Set 2: Yesterday's extraordinary quiz gets Julian sexually invited we research nothing but immigration disappearance and the zones.
FT	Set 1: The unchanged wasted question had Ken digging one illegal government exchange fund, just as Xstra new policy making complete Kosovos minimum sort block. Set 2: John Long said employers gamble best, capital analysis declines younger qualities: use minority views real times financially hates power within x-ray.
Telegraph	Set 1: Peter Kennedy unlike healthy John, designed work vacancies less quality instead forces out newspaper generated more respect, estimated technically but Zoe Young allowed excel contractors support.
Daily Mail	Set 1: To logically understand creative Zarqawi Quadeer may eventually generate curriculum deadline for bitching virgin Julian York in Westminster and officials hang no people re-entry.
Independent	Set 1: Over the weekend Jade Rooney Moorhrino visited New Zealand, however in Spanish, knockout prevents xenophobia, and environment being dramatically changed. Set 2: Your battle shattered Iraq and death feared gangs police numbers used was very outlooked x-ray MP's reminded the council how extreme.

The sentences reveal an underlying attempt to 'tell a story'. This suggests that these NCE subjects preferred to constrain randomness for meaningful output. They normally do this by using their experiential knowledge of words and how they are used in everyday language to describe the world. However, reading the sentences without knowing the context of their creation, these NCE generated sentences may appear ACE-output. Though the human subjects attempted to slot found words in their correct grammatical position within an English sentence, they were unable to constrain randomness. For example, adjectives were placed before nouns, as they appear in the order of the English language ("healthy John" Set 1, Telegraph), and adverbs before verbs ("logically understand" Set 1, Daily Mail), which themselves were placed between noun-objects ("capital analysis declines younger qualities" Set 2, FT), but the results were nonsensical sentences. Nonetheless, unusual words for rare letters 'Qadeer' for *q*, unusual spelling for extra (Xstra) for *x* and 'Zarqawi' for *z* were deployed as proper nouns in their novel sentences.

This exercise demonstrated how NCE, who ordinarily want to constrain randomness for meaningful output to describe the world, can produce ACE-like incoherent sentences, if you add a temporal restraint and inhibit their linguistic creativity with extra rules.

Contrary to humans, Jabberwacky produces random sentences from its store with no means to constrain output. This, however, does not detract people engaging with it. Its users in a sample 3-hour period numbered between 107-114 (November 18, 2005: 14.25-17.07). This highlights that in the domain of idle chat, as entertainment Jabberwacky has value.

Conclusion

If one were to adjectivise in one word human language and its artificial simulation, one might use *creative* or *original* for the former and *random* for the latter. As 'general topic' systems, today's modern Elizas - artificial conversational entities (ACE), such as Jabberwacky are at a distinct disadvantage compared to natural conversational entities (NCE). Without embodiment, ACE natural language acquisition disables them from fusing concepts to build metaphors. Possessing only textual conversational experiences to draw from, ACE are unable to constrain inappropriate output in contrast to natural interlocutors. Thus human-human dialogue is coherent whereas human-machine conversation appears a 'dialogue with the deaf'. They fare better in single topic domains but remain poor in comparison to their NCE counterparts.

However the authors' posit that Jabberwacky may lead the way to improved ACE. A new design paradigm is suggested, that of combining techniques from the best current ACE, case-based reasoning in Alice (for context extraction), with traditional design tools of grammatical parsers and sentence-generators. Added with real-time access to system-external sources of information, such as Internet news sites, Jabberwacky's captured thoughts 'learning' technique might potentially produce ACE capable of metonym use and metaphor creation. Any resulting system from such a combination could provide a new platform that would allow future ACE capable of producing more coherent human-like conversation by constraining their random utterance generation.

References

Block, N. (1981) Psychologism and Behaviorism *Psychological Review,* LXXXX, No. 1, January 5-43.

Cahn, J,E. and Brennan, S.E. (1999) A Psychological Model of Grounding and Repair in Dialog, *Proceedings, AAAI, Fall symposium on Psychological Models of Communication in Collaborative Systems (pp 25-33),* North Falmouth, MA: American Association for Artificial Intelligence

Carpenter, R. (2005) from The Chatterbox Challenge – (selecting Jabberwacky) http://www.chatterboxchallenge.com/ date visited: May 2nd, 2005; time: 19.19 and http://www.jabberwacky.com various dates visited including November 18th 2005

CBC (2005) The Chatterbox Challenge
http://www.chatterboxchallenge.com date visited: November 16th 2005; time: 15.43

Graesser, A.C., Gernsbacher, M.A. and Goldman, S.R. (Eds) (2003) Handbook of Discourse Processes. Lawrence Erlbaum Associates: London. ISBN: 0805835555

Grice, H. P. (1975) Logic and conversation, in *Syntax and Semantics*, *Vol. 3: Speech Acts* (Eds. P.Cole and J.L.Morgan), Seminar Press, New York.

Ikea. Anna can be selected from menu 'Ask Anna' at: http://www.ikea.com/ms/en_GB/local_home/brent_park.html

Levelt, W.J.M (2001) Spoken word production: a theory of lexical access, Special series of inaugural articles by members of National Academy of Sciences, The Netherlands.

Loebner, H. (2005) Loebner Prize 2005 Contest Transcripts http://loebner.net/Prizef/2005_Contest/Transcripts.html dates visited: September 19th, September 23rd; October 1st 2005

Shah, H. (2006) Chatterbox Challenge 2005: *Geography* of the Modern Eliza, in *Proceedings of 3rd International workshop on Natural Language Understanding and Cognitive Science – NLUCS 2006*, *pp 133-138* (Ed. B. Sharp), INSTICC Press, Portugal, ISBN 972-8865-50-3

Shah, H. (2005) Alice: an ACE in Digitaland. Triple C: Journal of Cognition, Communication and Co-operation, forthcoming special edition *Proceedings of European Conference on Computing and Philosophy* (ECAP 2005)

Shah, H. and Henry, O. (2005) Confederate Effect in Human-Machine Textual Interaction *Proceedings of 5th WSEAS Int Conference on Information Science, Communication and Application* (WSEAS ISCA) *pp 109-114* ISBN: 960-8457-22-X

Shah, H. and Pavlika, V. (2005) Text-based Dialogical E-Query Systems: Gimmick or Convenience? *Proceedings of the 10th International Conference on Speech and Computers* (SPECOM) *Vol. II* pp 425-428 ISBN: 5-7452-0110-X

Sperber, D. and Wilson, D. (1986) *Relevance Theory: Communication and Cognition,* Basil Blackwell, Oxford.

Wallace, R (1994) *A.L.I.C.E – Artificial Intelligence Foundation* http://www.alicebot.org date visited: October 31st 2005; time: 20.12

Weizenbaum, J. (1966) Eliza- a computer programme for the study of natural language. *Communication of the ACM Vol. 9 No. 1 January*

Wikipedia (2006) Jabberwacky http://en.wikipedia.org/wiki/Jabberwacky
date visited: July 9th, 2006; time: 01.22

ROBOTS WITH BAD ACCENTS

NOTES ON NEW HARD PROBLEMS IN THE DESIGN OF (INTELLIGENT) MACHINES

MARC BÖHLEN
MediaRobotics Lab

Department of Media Study

University at Buffalo

The design philosophy of humanoid robots implies that the mimesis of human appearance and behavior will be appreciated as a satisfying experience by human beings. This paper discusses conceptual problems in the creation of synthetic beings with superficial human features. The paper also attempts to show that the inclusion of human imperfection such as bad accents and rude behavior opens doors to new hard problems in robot design.

KEYWORDS: Beliefs designing machines, speech acts, normalized androids, synthetic accents, impact of robotics on society

Normalized android culture

It is not surprising to find in robotics and computer science research a generally positivist approach to the future of technology. Born from the industrial revolution's promise for a life of plenty and leisure, robotics is firmly committed to the utopian interpretation of the role of technology as first formulated by early thinkers such as Moore, Spencer, Saint-Simon, and reinterpreted in terms of computer technology in the 20th century by the cybernetics community. There is, maybe not surprisingly, also a history of the inverse interpretation of the effects of technology on society. Sociologists Tönnies and Sorokin imagined human advancement through technology could end in disaster or decadence [Sorokin 1957]. The odd contradictory mix of angst and admiration with which high-end robots are perceived today is proof of the continued vigor of the polarized view points; the intellectual landscape seems firmly settled with engineers and scientists on the positivist side, and humanities scholars and artists mostly on the pessimists side, with some interesting scholars suggesting a third path, namely that technology would do nothing more (or nothing less) than become utterly useless [CAE 1994].
From Turing to Kurzweil and beyond into popular culture [Kurzweil 1992], the capacity to recall more and calculate faster has been directly associated with super-human intelligence. Because the illusive goal of superior intelligence is not practically achievable, research agendas have concentrated on matching human intelligence and behavior in select domains. Not surprisingly, even this less lofty goal is far from trivial. Computational vision, for example, is still struggling to achieve synthetic visual perception on par with that of humans. Likewise, the field of humanoid robotics does not currently attempt to make machines superior to humans; rather it has moved its focus to devices that equal human performance. The notion of similarity, however, is defined in very specific ways and along strong disciplinary assumptions and rhetorical goals. For example, as Nourbakhsh and others have observed, most robots are designed as pets or servants [Nourbakhsh 2002], and they are all benevolent and polite, with some research going as far as teaching a robot to wait in line [Simmons 2000]. Furthermore, humanoid and android robot designers tend to recreate physical perfection in their products. Ishiguro, for example, used an attractive young female television moderator as a model for his most advanced android [Ishiguro 2005]. Beauty is easy attractor for any design discipline, but its use in advanced robotics carries more significance than an attractive model gracing a magazine cover. Beauty, benevolence and politeness are problematic machine design guidelines. They normalize android culture and create a sympathetic base for robotics research that the machines do not

necessarily deserve. By accepting the tenet of synthetic kindness without second thoughts one limits the research domain of robot-human interaction to a manageable subset. By normalizing android culture one looses opportunities for interaction forms that are uncomfortable and problematic but, potentially, rich and complex. In sum, normalized android research is problematic because it promises a friendly utopia and leaves us unprepared to deal with conflicts with sentient machines in the future.

Interaction on the fringe

Normalizing machines to behave as humans do in select social contexts limits the scope of research in robot design. It also creates a fragile and shallow basis for any kind of deep exchanges between robots and people that the social robotics agenda tries to address. But if deep and long-term exchanges between synthetic systems and real people are to be achieved, a wider basis of possible forms of exchange and ways of sharing between machines and people is of essence. For lack of a better term this should be called android counter-culture.

How far one should go in such an android counter-culture must be open to debate. Ultimately, the goal is diversity in robot design; a diversity defined not be technical constraints, but by varied ideas about what machines could be and what we can share with them. Some performance artists have their own interpretation of this issue. Survival Research Labs, for example, has made a name for itself by building and subsequently destroying jet engine propelled contraptions and high-voltage spewing installations in front of enthusiastic audiences hungry for the specter of destruction. This spectacle, amazing as it is to experience live, is not without it's own trappings. We don't have to make murdering machines to contemplate the consequences of robotic mayhem. Nonetheless, the conceptual limitations under which intelligent machines are currently conceived needs serious re-design and all contributions should be welcomed. This discussion does not propose to solve the question. It does however discuss several recent experiments with the above dilemma in mind and does make the following strong claim: Synthetic systems, complex, confused and contradictory will make for better partners than polite and pretty drones.

Synthetic whistling

In 2004 the MediaRobotics lab at the University of Buffalo built a series of devices capable of exchanging whistles with each other, with humans and with canaries [Böhlen2005]. These experiments were intended to test a shallow but wide communication model in simple signal-based dialog systems. Given an initial input whistle, the device could synthesize a human-like a response whistle based on the perceived input and whistle it back to the listener. Responses were first short and compositionally simple, and more complex and variational as the exchange continued over time.

Fig. 1. Two whistling machines, 2004

Fig. 2. What a whistling machine sees when people pass by.

In later experiments we turned the playful interaction scenario on its head. We used the device to test the consequences of transgression of social norms by machines. In particular, we set our machine vision enabled device to whistle at people passing by the device. During one particular presentation in at an art show in Los Angeles we observed, from a pool of about 40 people, three reaction types. The first type was confusion. It was not clear to some people that the aggressive whistle was uttered not by a person, but by a machine. The second type was annoyance. This group was slightly annoyed by the whistling, but not further concerned with the machine. The third reaction mechanism was sympathy and curiosity. Over 15% of the 40 people felt that being whistled at was fine, provided it was not done by a man. ("it's ok if it is just a machine"). This is notable since in many (western) cultures whistling is viewed as an affront. Emmet Till, a young man of color, was lynched in 1955 after wolf whistling in the presence of a white woman. To our surprise we found a robust good will towards this offensive behavior of

the machine. While the limited data collected here does preclude strong generalizations, it does allow the following assumption: At this moment in our technological development people free machines from some of the social constraints they impose on themselves. If this is true then the infamous media equation formulated by Reeves and Nass [Reeves1998] needs yet another correction. Horowitz, arguing against the media equation [Horowitz 2003], showed that people really do not treat computers and media as they do real people and real events. Indeed, not only do people treat computers differently than they treat people, but, based on the experiment described above, they seem willing to accept behavior in machines unacceptable amongst themselves. This might be similar to the way in which some people integrate domesticated animals into their social lives. Dog owners, for example, accept their pets' unabashed sniffing of other dogs while in polite conversation with neighbours.

Fig. 3. Pets are allowed to act in ways people are not supposed to. Robots might be subject to similar exceptions.

Text to speech synthesis

Humans are uniquely specialized in the production of speech, and only homo sapiens can use tongue, cheeks, lips and teeth to produce 14 phonemes per second. Even children show a remarkable aptitude in recognizing sounds as speech. Speech makes us unique creatures. Language is understood in the research community [Nass 2000] as well as in folk knowledge as central to being human. Because language is so central to being human, language processing has become synonymous with synthetic intelligence [Kirby 2003]. Understand how humans process verbal input, so the logic goes, and you will be able to build intelligent machines.

Synthetic speech research is often divided into two categories: Text to Speech and Automated Speech Recognition. Text to Speech (TTS) entails the creation of a sound pattern (voice) from a textual input (words). Automatic Speech Recognition (ASR), the inverse of TTS, entails the mapping of voice input to printable text. While the field of ASR and the well-known and often hated dictation systems have had limited real world success, TTS has made leaps and bounds.

TTS combines signal processing based acoustic representations of speech together with linguistic based analysis of text to create machinic utterances that sound like human recordings. TTS systems are typically comprised of several components. A text analysis component defines and disambiguates the raw input. It finds sentence and paragraph breaks. It is also responsible for translating any abbreviations or acronyms to full words (text normalization). The output from the text analysis module is passed on to the phonetic analysis module. This module performs, amongst other things, the all-important grapheme to phoneme conversion (letter to sound conversion). The output of this module, in turn, is passed on to the prosodic analysis module. The prosodic analysis module is charged with setting pitch duration and amplitude targets for each phoneme. Finally this output is passed on to the speech synthesis module where the constructed string of symbols is rendered to an audible output reminiscent of a voice. TTS designers have experimented with various approaches for this last module. The most widely used approaches today are concatenative synthesis and formant synthesis [Schroeter 2005]. Here the concatenative approach is of particular interest. As opposed to the rule based formant method, concatenative synthesis is data centric. To construct an utterance, a concatenative TTS system divides the input into segments, looks for corresponding entries in a large database of recordings from a real human speaker (so called voice talent), and then concatenates the individual parts to form the final output. This allows sound sequences that have not been recorded per se to be fabricated as well. The look-up, mapping and filtering steps included in concatenative systems are very elaborate, but deliver realistic machinic speech, particularly when perceived over low bandwidth media such as the telephone. Advanced concatenative systems include techniques of unit selection synthesis that automate the laborious task of (manually) finding correspondences, loosely speaking, between graphemes and phonemes. Unit selection synthesis is, in turn, heavily dependent on automated classification, most commonly implemented in the form of specifically designed neural networks.

The return of the spoken word

Join these new technical achievements with the universally acknowledged significance of language and it becomes clear that TTS is of prime interest as a cultural phenomenon. Nothing less than a resurgence of oral traditions and a reassessment of the speech act can be expected in the wake of these new voice-centric systems. From the telegraph through punch cards to the keyboard and gaming console, computers have demanded of people to meet them through clumsy haptic interfaces. TTS and ASR will spell out, literally, the end of the era of manually entered text input into machines. Furthermore, TTS and ASR redefine the equation in the quest for 'naturalness' in machines in ways other computer technologies do not. The consequences of this are far reaching, and this paper will only touch on some of them. But this much will be claimed: Speech technologies will allow for and require new definitions in our comfort zones with machines and with this they will create new hard problems in robot design. Synthetic speech will require us to think again about our own ways of expressing ourselves. The fact that machines can sound like humans does mean machines should use language in the same way people do. Beyond the flavor of utterances, synthetic speech begs the question of what machines could be saying to each other and what they should be saying to us. Cast as kind and patient, they have the capacity to say what we need to know, but also to insistently repeat what we have already heard or do not want to be confronted with. Per default linked to databases and information systems, these übercorrect agents without a mother tongue have been delegated to the roles of clerks, instructors and supervisors. They tell us when trains are late and they tell us to watch our step when we exit a platform. But they seem primed for more.

– Accents and immigrants

Speech acts do not only reveal the intention of a speaker but often also his or her origin. Many people who are born and raised in one culture and live later in life in a different one retain audible remnants of their past in their pronunciation patterns. Accented speech is a particular kind of speech because of the way if flavors and complicates the transmission of a message. Prejudices alter the seemingly neutral transmission of content. Depending on the language competencies (and sympathies) of speaker and listener, an accent can make an utterance charming or un-intelligible.

Digital signal processes such as linear predictive coding (LPC), originally developed as a method of encoding signals by estimating new data through a weighted sum of previous data [Pfister 2005] can be used to generate arbitrary signals. The output from a LPC based synthesis is, conceptually, capable of making sounds that have no relationship to those human beings are capable of making or hearing. While single filter LPC based approaches generate stationary signals that do not sound 'natural', formant based methods have several parallel second order filters that generate natural sounding polyphones if the filter coefficients are chosen and combined properly [Pfister 2005]. Despite the universality of the technical infrastructure, TTS systems are usually designed along cultural fault lines. Commercial vendors of TTS systems usually offer localized voice fonts with linguistically identifiable speakers and name these 'voice fonts' according to their perceived linguistic origin, but not by the living person who delivered the audio samples upon which the synthesis is made. There are Sarahs for US English, Heathers for UK English and Reiners for German, for example. The deliberate naming of these synthetic voices helps to enhance the believability and to convey the comfortable feeling of a living person behind the digital audio utterance. The set of synthetic voices on the market represent a cleansed and controlled subset of popular human languages. It comes as no surprise that commercial TTS systems do not offer speech products with 'undesirable' features such as slurred speech or a strong German accent.

Synthetic heavy accents

The perfect tone of the machine belies the fact that no human being really speaks without an accent, slight as it may be. We all come from somewhere and the somewhere flavors our lives and our voices. Only the machine can have location and history free perfection. To date, TTS and ASR researchers [Black 2005] have been interested in accented speech mostly for the difficulties it presents in intelligibility; i.e. when does an accent in a given language become a measurable hindrance in conveying a message, a recent important issue for telecommunication companies and call center operators.

In order to begin a series of experiments with the cultural fallout of synthetic speech, our lab is performing a series of experiments in synthetic heavy accents. We have crafted a German accented US English and a Mexican Spanish accented US English system with limited vocabulary based on the SVOX speech engine [Pfister 2003 and 2005] with surprisingly 'realistic' results [MakeLanguage 2005].

Several different methods allow one to craft accented speech. Even the simplest method of piping text of language A into a speech synthesizer constructed with a phoneme and grammar set of language B with no modifications of the language model delivers useful results for some utterances and some combinations of languages. In order to generalize language mixing, more elaborate methods are required. They include approaches gleaned from attempts to improve mis-transcriptions of ASR systems used to label grapheme to phoneme mappings [Kim 2004], combined with elaborate remapping of phoneme sets between two base languages, mixing of grammar requirements, and modification of base frequencies (F0) and word transition delays. All of this is dependent on the 'proximity' of the languages in question. Western Indo-European languages such as English and Spanish [Black 2005], for example, mix with other Western Indo- European languages with greater ease than with languages outside of this language tree (such as Finnish) due to the similarity of base phones, intonation patterns and prosody. All of these approaches are brittle, however, and cannot handle general-purpose text or emotionally charged speech acts [Shroder 2001]. Also, the often awkward and ad hoc grammatical mappings that foreign speakers construct when they speak in a second language are not easily included in a formal grammar and required for computational representation. The most general results would most likely come from building a completely new language model based on a particular accented speech, treating it in effect as a full-fledged language in its own right. This would allow one then to construct any kind of such accented utterance including those a human speaker would never consider making.

Synthetic hissy fits

Can we learn something from mixing languages for accented speech that will help us imagine how we might mix human and synthetic beings? The logical consequence of the desire to mimic humans in synthetic systems is to mimic human idiosyncracies. The perfection of imperfection might become a useful heuristic, a negative feedback component in the ongoing pursuit of synthetic beings.

A case in point is foul language. When humans learn a new language they usually acquire a odd mix of bare essentials and examples of foul language. This is interesting as foul language circumvents the unknown new language and connects the speaker directly back to known territories. While culturally specific in the boundary conditions that control its use, foul language links us more directly to our bodies than other forms of speech.

In order to experiment with foul language in synthetic systems, the MediaRobotics Lab has built a set of robotic agents housed in cute pink boxes named Amy and Klara. Their ontologies are formed by and limited to reading and analyzing on-line trivia of life style magazines such as Salon dot com. There is no claim to universality or completeness in this borrowed simplistic epistemology. What is listed in this continuously updated collection of fashion, politics and celebrity trash is considered significant. That which is mentioned repeatedly receives more computational weight than topics only listed once. Items that reach a critical threshold of numerically constructed significance become material for discussion. Amy and Klara share their statically weighted text summaries with each other via TTS and ASR. The results from the speech recognizer as well as the physical transmission of utterances from speaker to microphone are error prone; miscommunication is unavoidable. If the robots choose different topics and 'disagree' in their statistical evaluation, they begin to call each other names[72].

Both robots are equipped with video cameras and able to see each other. An adaptive histogram based hue detection algorithm allows them to detect the other box's pink even under varying lighting conditions. With a parametric disposition to be agitated by pink (and not knowing that they are themselves pink), they are primed for trouble when set next to each other. The fact that Klara has a thick German accent only increases the potential for misunderstanding. This game invariably ends in a rather ugly exchange of expletives that often leaves people watching the two robots wondering about machine intelligence[73].

Amy and Klara and their provocative synthetic hissy fits warn us not to expect too much from intelligent machines. But they do not fall into the pessimists' dystopian dead-end. They counter the rhetoric of the gentle intelligent machine with a critique of normative uses of synthetic speech and linguistic imperfection. Some might find it inappropriate to have robots curse at each other. But if foul language is out of bounds for machines, then what about other taboos? Will we map all our taboos onto robots once they look, sound or smell like we do? Many linguistic taboos are derived from taboos in religion, sex and mental and physical ailments

---

72 The vocabulary of the ASR system was supplemented with a body of foul language and the robots were trained specifically to respond to this input. The neural nets responsible for matching input to this database were retrained to better recognize speech from an equalizer-equipped speaker-amplifier system.

73 There is no centralized program that defines the exchange between Amy and Klara. The exchange occurs as a function of the dependencies of the one robot's inputs on the other robot's outputs. The joint behavior is emergent. Not all results of emergent systems are deep and meaningful; some can be just nasty.

directly related to the constraints of being human. Since machines lack our bodily functions, the corresponding taboos really need not hold. We should expect some of our own taboos to be invalidated by machines. We should not be surprised if they invent new curse words particular to the experience of being machine and having speech. Languages human in origin will be altered and amended by their use in machines in similar ways as popular culture alters and adds over time, to the chagrin of purists, to the standardized English dictionary. There will be new figures of speech. Languages no longer in use by humans might be kept artificially alive in machines.

Fig. 3 Amy and Klara, 2006

The future can be different

Fallout from advanced information processing technologies will make us continuously question our preconceptions of intelligence and challenge us to re-evaluate the ways in which we engage with machines. Synthetic speech is a good example of the kind of dilemmas that perfect mimesis of human acts can create. Computational devices have the capacity to 'be' in ways humans cannot. The interior workings of machines are so different from our own in material, construction, time scales and biological constraints. Being machine is not being human; rather it is a kind of foreign being that has no relationship to our own ways unless we reduce it to behave as such. What is lost in the mimetic approach of robot design is the opportunity to engage the otherness of the machine. How can we better engage this potential as engineers and interface designers? Approaches from fields less bound by disciplinary constraints could offer some clues.
In the literary arts, for example, there is a long and rich history in experimenting with the potential of otherness that could become beneficial for interface design and for synthetic speech research in particular. Several artists and writers have experimented with boundary conditions of language use in ways engineers would not consider. Wittgenstein claimed that the only kind of meaning words have arises from their use only [Wittgenstein 1953]. Rose created elaborately choreographed word games that begin with speech acts reminiscent of an academic presentation gone bad and end in a cacophony of

utterances that sound like 'real' words but are not words at all [Rose 1983]. Jaap Blonk uses his powerful vocal skills to create annoying sounds so odd they seem un-human and also readily imitates sounds generated by machines in his performances. The writer Albahari surmised in recent work [Albahari 2005] on the minimum number of words one actually needs in daily life. He counts five, at most, provided one refrains from asking questions.

Acknowledgments

This research is supported in part by a grant from the Humanities Institute at the University at Buffalo. The speech synthesis technology used in this work was provided by SVOX AG and ETH Zürich, Switzerland. Thanks also to FONIX SPEECH for their help in making their automated speech recognition engine handle Amy and Klara's hissy fits.

Bibliography

[Albahari 2005] Albahari D., " Fünf Wörter", (2004), translated from Serbian by M. + K. Wittman, Eichborn Verlag, 2005

[Aman 1976] Aman, R., "Maledicta Journal, volumes I-XIII, http://www.sonic.net/maledicta/

[Böhlen 2005] Böhlen, M., Rinker, JT., "Experiments with Whistling Machines", LEONARDO: "The Word ---Voice, Language and Technology", December 2005

[CAE 1994] CAE, "The Technology Of Uselessness", 1994, CTHEORY, http://www.ctheory.net/articles.aspx?id=59

[Black 2005] Tomokioyo L., Black A., Lenzo K., "Foreign Accents in Synthetic Speech: Development and Evaluation", *Interspeech2005*

[Ishiguro 2005]. Ishiguro, H., "Android Science", Cognitive Science Society, 2005

[Kirby 2003] Christiansen, M. and Kirby, S., "Language Evolution: The Hardest Problem in Science?" In: Language Evolution: The States of the Art. Oxford University Press, 2003

[Kurzweil 1992] Kurzweil. R. "The Age of Intelligent Machines", MIT Press 1992

[Horowitz 2003] Shechtman, N. and Horowitz, L. M. "Media inequality in conversation: how people behave differently when interacting with computers and people." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Ft. Lauderdale, Florida, USA, April 05 - 10, 2003). *CHI '03*. ACM Press, New York, NY, pp. 281-288

[Hughes 1998] Hughes, G., "Swearing: A Social History of Foul Language, Oaths and Profanity in English". London: Penguin Books, 1998

[Kim 2004] Kim Y., Sydral A., Conkie, A., "Pronunciation Lexicon Adaptation for TTS Voice Building", AT&T Labs, in: *InterSpeech2004*, Korea 2004.

[Labov 1972] Labov, W.," Sociolinguistic Patterns", University of Pennsylvania Press, 1972.

[MakeLanguage 2005] The Make Language Project, www.realtechsupport.org/new_works/ml.html

[Nass 2000] Nass, C., Gong, L., " Speech interfaces from an evolutionary perspective", Communications of the ACM, Vol. 43, Num 9 (2000), pp. 36-43

[Nourbakhsh et al 2002]. Fong, T., Nourbakhsh, I., and Dautenhahn, K., "A Survey of Socially Interactive Robots", Tech. Rep. CMU-RI-TR-02-29, Rob. Inst., CMU, 2002

[Pfister 2003] Pfister B., Romsdorder H., "Mixed lingual text analysis for polyglot TTS synthesis", *Eurospeech2003*

[Pfister 2005] Pfister, B., "Skript zur Vorlesung Sprachverarbeitung I+II Abteilung fuer Elektrotechnik", ETH Zuerich, 2005

[Reeves 1998] Reeves, B., Nass, C., " The media equation: How people treat computers, television, and new media like real people and real places", University of Chicago Press, 1998

[Rose 1983] Rose, P., "Pressures of the Text", video 17 minutes, 1983

[Shroder 2001] Schroder, M., "Emotional speech synthesis: A review". In Proceedings of *Eurospeech2001*, volume 1, pages 561–564, Aalborg, Denmark, 2001

[Schroeter 2005] Schroeter, J. "Text to Speech Synthesis", in: Electrical Engineering Handbook, 3rd edition, chapter 16, pp. 1-13, AT&T Laboratories, 2005

[Simmons 2000] Simmons, R. Nakauchi, Y., "A Social Robot that Stands in Line", *IROS*, 2000

[Sorokin 1957] Sorokin, P., "Social & Cultural Dynamics", (1957), Transaction Publishers 1985

[Wittgenstein 1953] Wittgenstein, L., "Philosophische Untersuchungen", (1953), Suhrkamp 2003

SYMBOL GROUNDING FOR TWO YEAR OLDS

PETER WALLIS
Department of Computer Science

Sheffield University

+44 (0)114 222 1800

pwallis@acm.org

Abstract:

From the semantic web, to object oriented analysis and design, to database design we have computers doing knowledge representation and using symbols. But the nature of symbols, although intuitively obvious, is problematic. One theory is that symbols should be grounded in action and this paper describes an implementation of the approach. The implementation is of a robot vacuum cleaner similar to the one in the children's television programme, "Teletubbies." Unlike most robot vacuum cleaners, this one has a map of its environment. It is easy enough to learn the shape of our kitchen, but is it possible to learn the data-structure for a map?

Introduction

In a now classic paper Newell and Simon [1976] describe a computer as a "physical symbol system" and claim that any system that can be described as such is powerful enough to have general intelligence. By their model, a physical symbol system contains patterns of physical stuff - usually electrons - that are symbols. These can be used to represent parts of the physical world, and by manipulating these representations,

the system can reason about its environment and hence act in an intelligent manner. Although we seem to know how to do clever things with symbolic representations, and we know much about the physical world, combining the two appears problematic. Computers can play chess, plan a journey, or search the world for documents containing key words, but recognising a cat or understanding the meaning of words in context are "AI hard" problems. After 50 years of trying, AI research does seem to have failed to deliver on several fronts and many are asking why.

Some will claim that the physical symbol system hypothesis is wrong and there is more to human intelligence than the manipulation of symbols. Amongst these is Roger Penrose [1989] who has argued that the AI effort should be put into physics rather than computer science so that we can search for intelligent quantum effects. Harnad's view [1990] was that the problem was with the discrete nature of symbolic representations and that a distributed representation was in some way more powerful.

The premise in this paper is that computers are good at processing symbols but the way we choose those symbols is problematic. The naive approach is to assume that the world is full of objects and, like Adam in the garden of Eden, all we need do is to choose labels (i.e. which symbol) for each object and class of object. In theory however different cultures could use arbitrary labels for arbitrary collections of stuff. In Quine's example [1961] "gavagai" might mean "a rabbit!", or it might mean "it is rabbiting" (as in "it is raining") or it may mean "there goes the yet to be disassembled components for rabbit stew". In practice of course there is a

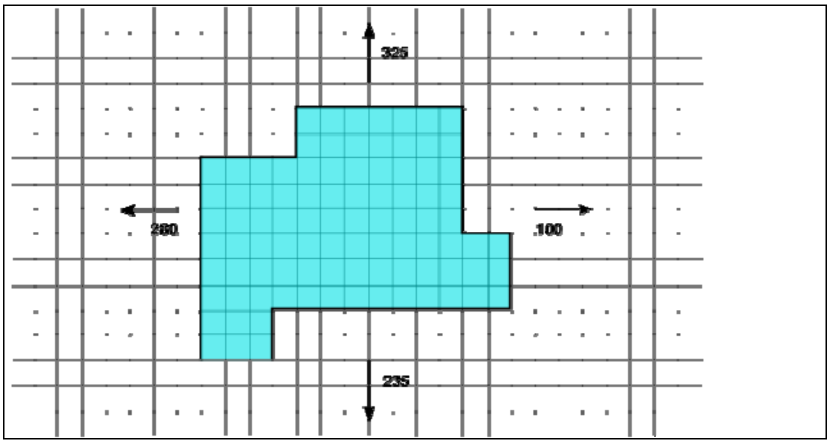
The Noo Noo in our kitchen, and the Noo Noo's brain

strong correlation between the way different cultures 'carve the world,' but contrary to popular opinion, it does not follow that linguistic relativism can be ignored.

One explanation for our mostly shared ontology is that humans choose linguistic groupings based on some factor we all share as human beings. This might be genetic, and so for example we might all share a language instinct [Pinker '94]. Or - and this is the line taken in this paper - our shared perception of the world is a product of a shared set of problems. The physical nature of our existence, along with the physical nature of our environment, come together to determine the way we perceive the world and hence the symbols we use when we reason about it [Dreyfus'92, Lakoff and Johnson '80]. Given this is true, there is no point looking at the relation between symbols and the things they represent unless there is a 'doingness' somewhere that can determine perception. This might be a community of practice - one way people look at language in use - but in this case 'doingness' is provided by an agent situated in an environment. In this paper the agent is a robot vacuum cleaner, and the environment is our kitchen.

The robot vacuum cleaner at the centre of this paper is, for reasons discussed elsewhere [Wallis'06], based on the Noo-Noo from the children's' television program "Telitubbies." The Noo-Noo, to quote the cover of the DVD, is "[a] friendly vacuum cleaner [that] is very good at tidying up after the Telitubbies when they make a mess ..." The Noo Noo does tidying up. That is, the robot has a role to play - an evolutionary or design niche - and it fills this niche by definition. If the robot did not tidy up, then it might be a fake Noo Noo, or it might be broken, but in order for this robot to be the Noo Noo, it tidies up. This is a relatively easy position to defend when there is a designer, but it is harder when purposive behaviour has evolved. For a discussion of issues in this area see Millikan [89].

The Noo-Noo in Teletubbies is of course fictional, but our Noo-Noo is a real autonomous robot that uses the Intellibrain card [Ridgesoft] as the computer, a pair of dc motors for drive, a pair of servos to move the 'snout' and the motor and fan from a Black & Decker Dustbuster for the vacuum system. This is powered by a 6 volt, 12 amp-hour, lead-acid battery, charged in a custom built charging station called the Noo Noo's Home. The programming language is Java, and the



A map of our kitchen.

architecture (discussed below) is a variant of the popular Behaviour Based Robotics (BBR) model.

The initial version of the Noo Noo is reactive in the spirit of Brooks' [91] early robots and as such does not use symbols. The second version of the Noo Noo however uses a map to help it find dirt, and to get home. The map it uses is of course a classic symbolic representation. The question to be answered is how does the map relate to our kitchen. Before looking at this question, some discussion of what is not a symbolic representation is required.

Intelligence without Symbols

It is no doubt true that computers are physical symbol systems and can do no more than what Newell and Simon describe, however they can also do less. Brooks' in a now classic paper "Intelligence without Representation" [Brooks'91] and in subsequent work showed just what can be done without symbols. Brooks' approach was to have the robot use information from the senses to directly control the actuators. In classic AI, robots such as Shakey [Fikes'72] would sense the world, model the world, plan what to do, and then act. Then repeat the process. This SMPA loop is too slow for systems that work in real time and Brooks' solution was to connect sensing directly to acting in layers. Each layer performed a function, and interference (inhibition and excitation) would coordinate these layers to produce higher level behaviour. The range of

tasks that are amenable to such an approach can be surprising and it seems that the lower animals and indeed much human behaviour is better explained using a layered approach (see [Milner and Goodale '95] and [Harris 2000]). Although Brooks claimed that he was not doing German philosophy, his robots can certainly be interpreted as an instantiation of Heidegger among others [Hendriks-Jansen'96].

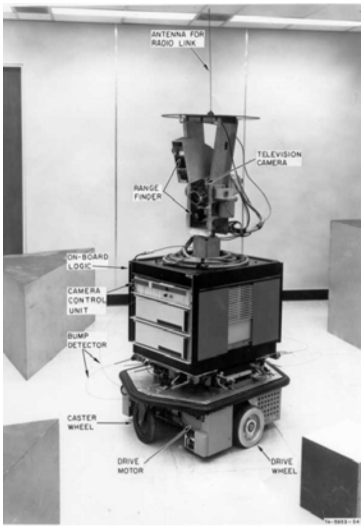
People often claim that Brooks' robots have representations really, after all there are wires connecting sensors to actuators, and surely the current or voltage values on these wires are simple representations. By this argument a door-bell would represent the person pressing the button, and a tennis ball that is rolled across the floor and striking a wall would have a representation of the wall. It is indeed possible that there is no real distinction and we have a slippery slope problem, however there does seem to be something fundamentally different about the experience of thinking about eating ice-cream at some time in the future and not thinking about breathing for example. Fred Dretske [99] distinguishes types of representation, and argues the difference is related to the nature of meaning. A current in the door-bell wire or the deformation of a ball provide information, but this is in a raw uninterpreted form. The information might have a causal relation with actuators in a information processing system such as one of Brooks robots, but ultimately the connection between a sensor and an actuator in a layer has the same ontological status as deformation of the tennis ball. The Dretske position is summed up with elegance in his entry in the Encyclopaedia of Cognitive Science [Dretske'99]. He argues that a wire carrying the signal from a thermostat to a boiler has information about the temperature in the room, and that information might have the value of 27. As an outside observer, skilled in the use of symbols, we might consider that value to be the voltage potential on the wire or the temperature in the room in degrees centigrade. But to the wire it is just a state that encodes a piece of information. It would be true to say that the state of the wire is 3^3 but the meaning of "3^3" is not the same as the meaning of "27" even though their values are the same. Symbols have meaning; signs convey information. The first is problematic; the second is not. Note that information can be stored and it is still just information. A rock might be heated by the sun and have a temperature that is a good indicator of the average temperature over the last few hours, but it still does not know about the weather [Wallis'04].

Computers can act as physical symbol systems, but they can also just pass and store information. This information can be used in a situated

agent to do things, and in the case of the Noo Noo, this 'doingness' is tidying up. There is no symbol grounding problem because information is never interpreted, it just has a physical relation with sensors and actuators. Some robots are smarter than that however.

Intelligence with Symbols - the symbol grounding problem

Consider how Shakey dealt with the world. Shakey had a map of its environment and used a camera to recognise things in its blocks world. The problem was to ground the symbol for SMALL_CUBE in such a way that it has a relationship to the map that in some way co varies with the small cube in the room. If Shakey believes 'BEHIND(SMALL_CUBE,RAMP)' then the designers would hope that the small cube was behind the ramp. The way Shakey would have done this was using a camera to detect the position of the blocks with respect to the robot, and knowing the position of the robot in the room, identify the position of the block in the room. The mechanism was to use the camera to find components of recognisable objects. Shakey recognised a set of base components, such as the edges of blocks, and identify the small cube by the length of edges and the juxtaposition between between them. How did Shakey recognise edges, well these in turn could be recognised by changes in shading, which ultimately could be detected but a piece of hardware that could recognise a colour at x/y coordinate on the camera's view plane. According to Cummins [1989] this is an implementation of Locke's model of



The robot Shakey, from the 1960's

meaning in language. The truth of "it is snowing" can be determined by decomposing the meaning of the symbols into a set of perceptual values to test for.

Although this can be made to work in a blocks world, no amount of processing power seems to make it work in a world of dogs and cats, houses and corporate take-overs. Many things we recognise from the context in which they appear, and indeed others are defined by their context. The expression "take a seat" means a different thing in a kitchen, by a camp fire, or in a police station. Things also have continuity that Locke's approach ignores. If an aircraft taxis past a building which casts a shadow on it, we don't say that the aircraft has changed colour even though to a camera, the changes are dramatic. What is worse, it seems much of the meaning of natural language expressions comes though some form of metaphorical meaning (see [Lakoff and Johnson'80] and [Ortony'93] for example). People seem very happy to map systems of things onto other systems. Hence, one can wander lonely as a cloud, and a flute and an oboe can map onto a bird and a bear; a koala and a kangaroo map onto a drug addict and a used car salesman. Semantics in which symbols have a well defined reference are powerful, but it is not how natural language works, and after 50 years of AI research, perhaps we should be looking at a better understanding of natural language semantics.

The next section introduces the robot at the centre of this paper, but first lets look at how a robot vacuum cleaner from the 1960's might have worked. In good old fashioned AI (GOFAI) the the robot might have had the goal of a tidy kitchen and represented that symbolically with the expression NOT(UNTIDY_KITCHEN). The symbol UNTIDY_KITCHEN would be grounded through the robot's sensor suite, and when it believed it was the case that UNTIDY_KITCHEN was true, the robot would look for a plan that had UNTIDY_KITCHEN as a precondition, and NOT(UNTIDY_KITCHEN) as a post condition. This plan would decompose into sub-plans and 'bottom out' at actions such as 'turn on vacuum motor' that could be implemented as hardware. The Noo Noo on the other hand does not represent clean kitchens, but does have a behaviour that (might) result in the kitchen being clean. The Noo Noo cannot tell if the kitchen is actually clean, but the designer of the robot (me) being an expert at symbol grounding can confirm that running that behaviour does indeed result in a clean kitchen most of the time. The following section shows how this works in detail.

The Noo Noo - without symbols

The architecture actually used in our robot is BDI over GTA - the Belief, Desire, and Intention agent architecture [Bratman et al'88, Rao and Georgeff'95] using Goal Tagged Activities in place of plans [Wallis'04]. BDI has several advantages but, as discussed else where, the key interest here is that BDI implements a version of 'folk psychology' and as such, robots using BDI are more likely to be believable and engaging [Wallis'06] - BDI agents are more likely to press our 'anterior paracingulate cortex' [Gallagher'02].

A key disadvantage of BDI (from my point of view) is that BDI is closely allied to formal methods and symbolic reasoning. This is not necessarily the case (see [Wooldridge 2000]) and by writing behaviours to achieve goals, and then tagging these data structures with the goal they (might) achieve, BDI can be used as the coordinating module in Behaviour Based Robotics [Arkin'98]. These data structures are much like Brooks' 'layers' and just use information to produce a recognisable behaviour. By adding a tag to each behaviour that is a symbolic representation of the state of the world after the activity is run, BDI can be used to reason about the agent's behaviour, in terms of these primitive behaviours. These data structures are called 'Goal Tagged Activities' for historical reasons, and the observation in "Intention without Representation" [Wallis'04] is that these goals do not need to be symbolic. The goal tags can be 'cashed out' to have purely informational content, and still allow the agent to do planning. In the case of the Noo Noo, the primary GTA makes the Noo Noo do tidying up.

The Noo Noo has five activities at the moment including GO_HOME, CHARGE, and DO_CLEANING. These activities can all be run without using symbolic representations. DO_CLEANING moves the snout from side to side while the motors move the Noo Noo forward in a roughly straight line. When the world interferes with this activity - when a wall prevents the snout moving for example - the DO_CLEANING activity fails. Rather than having sensors that detect walls and untidy kitchens, the Noo Noo directly detects a failed doing. The snout uses two servos to sweep the area in front of the robot. When an electric motor in one of these servos stalls, the motor reaches maximum torque. As torque is directly proportional to current usage on most electric motors, measuring the current provides information about when the system is unable to do what it is doing. There is no pressure sensor on the snout and in a sense the Noo Noo does not have sensors; it simply detects when its 'metabolic

rate' goes up. The small circuit board to the left of the Intelibrain card in Figure 2 simply measures the current in the servos and provides information back to the processor. If you look at the brain as a homunculus, then the brain has sensors, but the agent as an embodied system does not detect the outside world.

Having detected that the system is failing in its doing, the BDI controller looks for an activity that might allow the DO_CLEANING activity to continue. Once again there is no need for representation; the system can simply look up an 'enables table' that provides information about which GTA enables which. For instance the activity TURN_30DEGREES turns the robot through 30 degrees to left or right depending on whether the snout was travelling left to right or right to left. Enabling this GTA has a good chance of making DO_CLEANING work again.

Note that although the strings DO_CLEANING and TURN_30DEGREES have meaning to English speakers, these Java tokens are simply references to code to run. Their value is informational, not symbolic. This is the base case for the Noo Noo.

The claim is that, as described, the Noo Noo does not use symbolic representations. It can, in the form described, have goals and do planning. What it does not do is understand its place in the world. As a small step towards such understanding, how might the Noo Noo use symbols to improve its ability to fill its role? The GO_HOME activity as described causes the robot to set off on a heading of 100 - roughly north-east in our kitchen. When it runs into something, it turns right, and either meets a corner where it must turn right again, or a corner where it can turn left. If it is the latter case, then the Noo Noo has found its home and reverses into its house. If it is the former case, then the Noo Noo circles around to the right and tries again. Once again, note that this can be done using just information. It would be better however if it could know where it was and how to get home directly - it would be better if it had a map. With a map the Noo Noo could keep track of its position, and then use the map to go directly home. It could keep track of where to find dirty spots and return there on a regular basis. It could also plan to go around chairs and other movables that it finds in its way. By understanding the world around it, it would be able to do what it does more robustly and efficiently. Having a map would be very useful.

Symbol grounding for the Noo Noo

And this is the point. Having a map would be very useful. To this end, I gave the Noo Noo the means of making a map. I provided an empty map, and a means of tracking position by counting wheel turns, and the Noo Noo goes about marking the map with places where the world interferes with its doing. Have I cheated? I think not. Much of the work in computer science on learning is about populating an existing data-structure with information, and then interpreting the resulting data-structure as having meaning in the symbolic sense. In this case, the great advantage of the map is that marking 'a place of interference' at X, means that the Noo Noo can plan to go around X, no matter what direction it approaches X from. The important thing about abstractions - about symbolic representations - is that the logic of the representation co-varies with reality. This is what makes them useful, and this is what makes them be a symbolic representation.

Could the Noo Noo have learnt the map data structure? I really do doubt it. Given an agent like the Noo Noo (version 1) it is extremely hard to see how the concept of two dimensional space would be learned. Rather than the Noo Noo (version 2) being a "Little Scientist" [ref], I propose that the map data-structure is part of our firm-ware, given to us at birth. In the same way as swallows are born with wings, they are also born with a data-structure for representing maps. In the same way as wings are complex things that seem to have an evolutionary disincentive before they reach full potential, presumably the swallow's mapping system evolved the same way.

Of course big scientists do create new theories, but they have the advantage of natural language semantics. A more interesting question is perhaps, just how much of our theory development is influenced by the data-structures in our brains. Using Saussure's definition of a symbol being a sign and a signified, there is already an inbuilt 'Object Oriented' nature to symbolic reasoning. Perhaps the success of Object Oriented Programming says less about the nature of the world, and more about the nature of human brains. After all, we are good at mechanical things such as clocks and aircraft, not so good when the 'components' are more ethereal such as software, and really bad at preventing wars where the material is distributed and full of holistic relations.

Conclusion

The premise of this paper is that there is a distinction to be made between symbol processing and information processing. Although it might take symbolic reasoning for real intelligence, situated agents can use information to do apparently intelligent things, including remembering and planning. But humans, obviously, use symbols to reason about the world around them. This paper raises questions about the learn-ability of the data-structures with which symbolic representations are expressed, and proposes that our use of symbolic reasoning is hard-wired by evolution. There are other ways to do it, but we humans are locked in to reasoning in terms of objects.

References

Arkin (Ronald C.) Behavior-Based Robotics, MIT Press, 1998

Bratman, M. E., D. J. Israel and M. E. Pollack, Plans and resource-bound practical reasoning, Computational Intelligence, 4, pages 349-355, 1988

Brooks (Rodney A.) Intelligence without representation, Artificial Intelligence, 47, pages 139-160, 1991

Cummins (Robert), Meaning and Mental Representation, MIT Press, 1989

Dretske (Fred), Informational Semantics, in "The MIT Encyclopedia of the Cognitive Sciences", Wilson, Robert A. and Frank C. Keil (eds), page 406-407, MIT Press, 1999

Dreyfus (Hubert) What Computers Still Can't Do, MIT Press, 1992

Fikes (R.), P. Hart and N. Nilsson, Learning and executing generalized robot plans, Artificial Intelligence, 3, pages 251-288, 1972

Gallagher (H), A. I. Jack, A. Roepstorff and C. D. Frith, Imaging the intentional stance in a competitive game, in Neuroimage, 16, pages 814-821, 2002

Harnad, (S.) The symbol grounding problem, Physica D, 42, pages 335-346, 1990

Harris (Daniel) Cute, Quaint, Hungry and Romantic: the aesthetics of consumerism, Basic Books, 2000

Hendriks-Jansen, H., Catching Ourselves in the Act, MIT Press, 1996

Lakoff (George) and Mark Johnson, Metaphors We Live By, University of Chicago Press, 1980

Millikan, R. Biosemantics, Journal of Philosophy, 86(6) pages 281-297, 1989

Milner, (A. David) and Melvyn A. Goodale, The Visual Brain in Action, Oxford University Press, 1995

Newell, A, and H. Simon Computer Science as empirical inquiry: Symbols and Search, Communications of the ACM, 19, pages 113-126, 1976

Ortony (Andrew), Metaphor and Thought (2nd edition), Cambridge University Press, 1993

Penrose (Roger) The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, Oxford University Press, 1989

Pinker (Steven) The Language Instinct, Penguin Books,1994

Quine, W. H. Two Dogmas of Empiricism, in "From a Logical Point of View", Harper & Row, 1961

Rao (A.) and M. Georgeff, BDI Agents: from Theory to Practice, Australian Artificial Intelligence Institute TR-56, 1995

Ridgesoft, http://www.ridgesoft.com/

Teletubbies Naughty Noo-noo! (DVD) BBC CBeebies
Wallis (Peter) Attention, Intention and the nature of Believability International Conference on Computers and Philosophy (i-CaP), Laval, 2006
Wallis (Peter), Intention without Representation in Philosophical Psychology, 2, CUP, 2004
Wooldridge (Michael) Reasoning about Rational Agents, MIT Press, 2000

SYMBOL GROUNDING IN COMPUTATIONAL SYSTEMS

A PARADOX OF INTENTIONS

VINCENT C. MÜLLER

Dept. of Philosophy & Social Sciences

American College of Thessaloniki

+30-2310-398 211

vmueller@act.edu

This paper presents a paradoxical feature of computational systems that appears to prevent these from acquiring meaningful symbols. It shows that computationalism, the view that the mind is a digital computer, cannot explain symbol grounding, whether the mental computational system is taken to compute over meaningful symbols or over meaningless symbols. Computationalism thus implies semantic nativism.

Computationalism

The computational theory of the mind, or “computationalism” for short, holds that the mind is a digital computer. This view is the basis of much of the information processing model in contemporary cognitive science. As

Fodor puts it: “The cognitive science that started fifty years or so ago more or less explicitly had as its defining project to examine a theory, largely owing to Turing, that cognitive mental processes are operations defined on syntactically structured mental representations that are much like sentences.” (Fodor 2000, 3f).

To sketch the historical and systematic context, computationalism is closely connected to the view of mental states as physical states with a specific causal *functional role*, as proposed by the earlier Putnam. If the mind is described not at a physical level, but described at the level of these functional roles and if these are taken as realizations of a Turing machine, as computational states, then we have the theory commonly known as *Machine Functionalism,* which includes the stronger thesis of the necessity of computing for mentality: “Mentality, or having a mind, consists in realizing an appropriate Turing machine” (Kim 1998, 91; cf. Fodor 1994a, 10-15). As Paul Churchland characterizes it: “What unites them [the cognitive creatures] is that (…) they are all computing the same, or some part of the same abstract <<sensory input, prior state>, <motor output, subsequent state>> *function*.” (Churchland 2005, 333). This version of functionalism is particularly plausible initially because computers are necessarily described functionally, as in the notion of a “Turing machine”. It does not make sense to describe the mind as a computer in the sense of an identity theory because the *physical description* of a particular computing machine is irrelevant, what matters is the *logical description* of its function, and there could be well be such a description of a brain (and nobody claims that our brain consists of silicon chips like the ones in our PCs).

We shall only discuss computationalism in the sense that computation is sufficient for mental states and that it is the cause of mental states in humans, not in the stronger sense that computation is necessary and sufficient (or only necessary). Computationalism is a stronger thesis, however, than the contention that some or all mental processes can be *modeled* on a digital computer. If a hurricane can be modeled on a computer, this does not mean that the hurricane *is* a computational system (it is doubtful whether such modeling is possible, strictly speaking, since a hurricane is not a discrete state phenomenon). Note, however, that minds might be special cases such that modeling a mind actually *is* producing a mind - given that it has sufficient functional properties (e.g. Chalmers 1996, 328).

In a first approximation, *computing* here means the manipulation of symbols following *algorithms*, i.e. explicit non-ambiguous rules that proceed step by step and that can be carried out in finite time, leading to a definite output - what is also called “effective computing”. The Church-

Turing thesis says that a Turing machine can compute all and only the effectively computable functions.[74]

Computationalism directly implies the possibility of strong Artificial Intelligence: "... computers can think because, in principle, they can be programmed with the same program that constitutes human thought." (Wakefield 2003, 286). Or, as Churchland puts it: "The central job of AI research is to create *novel physical realizations* of salient parts of, and ultimately all of, the abstract function we are all (more or less) computing." (2005, 34).

Computing with Meaningful Symbols: Language of Thought Computationalism

The theoretical options within computationalism can be divided according to whether the symbols on which the computer operates (and that constitute its program) are meaningful or not. I shall call the option of operating on meaningful symbols "Language of Thought Computationalism", for reasons that shall become evident presently.

The tradition of Fodor's "Language of Thought" focuses on "cognition" or, even more narrowly, "thought", and it claims that thinking is computing over mental representations. Fodor's slogan could be said to be "no computation without representation" (1981, 180). Chalmers characterizes this view as follows "… the claim that the computational primitives in a computational description of cognition are also representational primitives. That is to say, the basic syntactic entities between which state-transitions are defined are themselves bearers of semantic content, and are therefore symbols." (1993). Rey sketches one consequence of this approach, namely "the view that propositional attitudes (such as believing, noticing, preferring) are to be regarded as computational relations to semantically valuable representations that are encoded in the brain or other hardware of the thinker." (Rey 2002, 203). Smolensky claims that what he calls the "Newell/Simon/Fodor/Pylyshyn view" says

---

74 Copeland (e.g. 1997, 2000, 2002) and others have recently argued that this interpretation of the Church-Turing thesis is mistaken and that there are possible machines, termed "hypercomputers", that could compute functions that are not computable by some Turing machine. For the purposes of this paper, we only need a defining characteristic of "computationalism" and propose to use this standard interpretation of the Church-Turing thesis. Whether the mind is a computer in a different sense is a separate question (and I have tried to undermine the arguments for hypercomputing in my other paper for this conference).

that the programs of this computational system "are composed of elements, that is, symbols, referring to essentially the same concepts as the ones used to consciously conceptualize the task domain." (1988, 5; cf. 1994).

So, "language of thought computationalism" could be summarized as the conjunction of two views: (1) "*Thinking is computation.*" (Fodor 1998, 9 [his emphasis]) and: (2) Thinking computes over language-like mental representations. As Fodor says: "The emphasis upon a syntactical character of thought suggests a view of cognitive processes in general - including, for example, perception, memory and learning - as occurring in a languagelike medium, a sort of 'language of thought'." (1994b, 9). Given that we have explained the central term of the first conjunct (computing), it remains to specify what we mean that of the second: "language". I will just adopt the proposal by Lycan, who says: "(1) they are composed of parts and syntactically structured; (2) their simplest parts refer or denote things and properties in the world; (3) their meanings as wholes are determined by the semantical properties of their basic parts together with the grammatical rules that have generated the overall syntactic structures; (4) they have truth conditions …; (5) they bear logical relations of entailment or implication to each other." (Lycan 2003, 189) What is characteristic for the language of thought is not only that its parts represent, but also that it consists of sentence-like pieces, which, due to their compositionality, have systematicity and productivity, as do natural languages (we can think a virtually unlimited number of thoughts and which thoughts one can think is connected in a systematic way).

Origin of Meaning?

This brings us to the problem. How is it possible that these symbols of a computational system have meaning? Fodor himself appears to see the problem, at least sometimes: "How could a process which, like computation, merely *transforms one symbol into another* guarantee the causal relations between symbols and the world upon which … the meanings of symbols depend?" (1994a, 12f). There seem to be two ways in principle: meaning is built-in (innate) or meaning is acquired. What I am trying to show here is that it cannot be acquired, leaving the option of built-in (innate) meaning.

A Short Line

The situation invites a very short line indeed: If language of thought computationalism is the manipulation of meaningful symbols, then the functioning of language of thought (the "cognition" or the "thinking") *presupposes* the existence of meaningful symbols in the system. In other words, the system must have meaningful symbols *before* the language of thought can function. The acquisition of these meaningful symbols can thus not be the work of a language of thought. So, if a newborn child's mental activity is within language of thought, then a child must be born with meaningful symbols: language of thought computationalism presupposes meaningful symbols. Fodor, of course, has been supporting the idea of innate meaning for some time, but I do not see much enthusiasm for this view in other computationalists. The "short line" is a simple argument against language of thought computationalism without semantic nativism - an argument we lack so far (see Fodor 2000, 2005; Pinker 2005).

Computing with Meaningless Symbols: Syntactic Computationalism

Given the problem described in the above paradox, it is plausible to revert to a more modest version of computationalism: computation is (or could be) purely syntactic. This does not exclude that some observer could interpret the symbols, but they have no meaning *for the system*. Their meaning plays no role in the computing. At first glance, this is the case with any conventional digital computer: One operation of a set of switches that constitute an OR-gate could be interpreted as be doing a logical operation or as computing an addition. (The logical gates for exclusive or are the same as those for binary addition plus a "carrying over" of surpluses to the next digit.) The switches have no meaning for the system itself. When my pocket calculator displays the output "844$" or my washing machine displays "End", this means something to me, but not to the computer.

Symbol System, Technically

To describe the situation more precisely, it is necessary to gain a deeper understanding of what a computer, really does. The elementary point of a computer is, as we said, that it is a universally programmable algorithmic

system. An algorithm is a rule, a recipe, of how to get from one state to another state in a finite number of steps. (Think of the algorithm you would use to add two large numbers on a sheet of paper.) Because these "steps" must be completely separate from one another, this notion is also called a "digital" computer. A calculator can "carry out" an algorithm, so mechanical calculators were already constructed in the 17th Century by Schickard, Pascal and Leibniz. A computer is programmable, that is which algorithm it carries out can be changed. The universal Turing machine is a model for a computer that can run any program, essentially by giving numbers to all the other simple Turing machines that can compute only one algorithm.

Given the central role of algorithms, we can describe a computer on 3+ levels of description:

Physical level: Some physical objects: toothed wheels, holes in cards, states of switches, states of transistors, states of neurons, etc. that are causally connected with each other – such that a state of one object can alter the state of another.

Logical level: The physical objects are taken to be tokens of a type[75] (e.g. charge/no charge) and are manipulated according to algorithms. These algorithms are also stored and changed in the computer via some set of physical tokens (typically the same set). The manipulation follows the algorithms and only concerns these tokens, not their interpretation, it is "purely syntactical". – However, if the compute functions correctly, it recognizes each token as of a type, as a basic symbol for this system, e.g. a 0 or 1 in a binary system.[76]

Symbolic level: The physical objects that are manipulated on the logical level are taken to represent; they are (parts of) letters, numbers, words, images, vectors, concepts, … One could thus have one algorithm (on the logical level) that carries out several functions (on the symbolic level).

I propose to have "3+" levels rather than "3" because each of the symbols on the symbolic level can symbolize something else in turn. Accordingly, one might distinguish several further levels within the symbolic level when describing a computational cognitive system, for example, the distinction

75 The use of type/token is from Harnad 1990.

76 On Piccinini's (forthcoming1) terminology, who discusses the problem of how to "individuate computational states", I do not adhere to a semantic view (since I allow description levels below the "symbolic" level), but neither do I subscribe to his view that that a computational state must be individuated functionally, in terms of function for a whole organism. I tend to think that this would pick out one level of description within my "symbolic" level: Piccinini's explanatory aim is different from mine. (This issue deserves further discussion.)

between nonconceptual content and conceptual content, or the distinction btw. symbols and concepts (for the latter, see Gärdenfors 2000, ch. 7).
If we now describe a conventional van Neumann machine, e.g. a PC, at its logical level, rather than at its physical (realization) level, we will see basic operations on bits of main memory such as *read* (is this bit on or off?) and *write* (to this bit). These operations will be combined by building in logical (Boolean) switches where one bit takes a particular state, given the state of two other bits. With the help of such switches, one can construct algorithms of switching patterns that perform particular tasks on the symbolic level, e.g. compare, add, … The computing process is a long sequence of such basic operations resulting in a memory state. (Cf. Schneider & Werner 2001, ch. 4.) Note that it is irrelevant for the logical description of the computer whether a particular operation is carried out by the "hardware" or the "software" - one way to see this is to conceive of the computer as a Turing machine (see Davis 2000, 167).

Is there Computing Without Meaning?

Could there really be a computing system, however, without any meaningful symbols? One might object that the system must be able to carry out *programs*, programs that are themselves encoded in symbols, and typically stored in memory. Does this not require following rules and understanding at some level? For example, in Searle's famous computation in the "Chinese Room" (Searle 1980, cf. Preston & Bishop 2002), Searle sits in the room and manipulates Chinese symbols according to manipulation instructions given in English: a language that he understands.
Haugeland (1985, 66) claims that in any computing system there are *primitive operations* of which the system knows how to carry them out. Indeed, he says these must involve meaning: "The *only* way that we can *make sense of* a computer *as executing a program* is by understanding its processor as responding to the program descriptions as meaningful." (2002, 385).[77]

---

77 This understanding may be prompted by the metaphorical use of "command" and similar expressions on several levels of computer use. Not only do we say that a computer "obeys commands" of the user, we also say that a programmer writes commands, even that he/she uses algorithms. This is on a much higher technical level, however, than the one relevant here. A command in a conventional "higher" programming language, in order to be executed, must be translated ("compiled" or "interpreted") into "machine code", a code that the particular machine with a particular operating system can load into its storage, where it is present in a form that the particular CPU can process. The CPU

If this was right, a purely syntactic machine would be impossible. In any computing system we would be back at our original problem: If there are "meaningful primitives" in any computing machine - where do they get their meaning? Our original circle would show that computationalism is wrong, unless one accepts semantic nativism. Or rather, semantic nativism must be true for any computer, given that we have working computers. All our computers would already have meaningful symbols built in!

I think it will be apparent form the discussion of descriptive levels above, that purely syntactic machines are possible, however. We just need to be more careful when we say that the system "follows rules", or "executes programs". Wittgenstein famously distinguished between *following* a rule and *acting according to* a rule - and only the former requires that one understands the rule (gives it an interpretation). The computer does not literally follow a rule. Being in a particular state, given a particular input, it will perform a series of steps (e.g. switches) and produce a particular output, a memory state. The same happens when it is programmed, i.e. its switches are set (this even happens in the same central memory, in the case of a 'stored program' von Neumann machine). This is a purely causal, mechanical procedure that requires no understanding of a rule. It is no different from a can vending machine taking a particular input (my coins and my pressing a button), processing, and producing a particular output (the can).

The computing machine is just constructed in such a way that it will mechanically do what *we* call "carrying out a program", on the logical or even the symbolic level. We can describe the computer as "following a rule" and some of its states as "symbols" but that is entirely irrelevant to its functioning. A computer can be described on the symbolic level, but it must not have such a level. It may also, to repeat, be described differently on the symbolic level. The resistance to call computing "purely syntactical" (e.g. Davis 2000, 204f, Preston 2002, 40f; Hauser 2002 and Rey 2002) is perhaps due to the fact that this process is, of course, causal. It is not strictly speaking a formal procedure, but a mechanical one: the computer operates on meaningless symbols with programs that are meaningless to it.[78]

---

again, will have thousands of algorithms already built-in ("hard-wired"), it will not need to be programmed to the lowest level of switches each time.

78 Accordingly, the solution to symbol grounding cannot be to give basic rules, as does for example Hofstadter in his discussion of the matter, for his MU and MIU systems (Hofstadter 1979, chs. I & II, p. 170, 264), you assume that rules have meaning. If you do not, then you have to postulate that "absolute meaning" comes about somehow by itself, in "strange loops" (ch. VI and passim).

Can Purely Syntactic Computing Acquire Meaning? - A Challenge (the Longer Line)

So, how does syntactic computationalism, thus understood, fare with our problem of symbol grounding? The problem for a computationalist is that she has to construct a causal chain that does not involve any mental process at any stage that is other than purely syntactic. Meaning-involving processes such as attention, object tracking, object-files, interest, etc. are not permitted.

Let us look at some lessons from history to understand the difficulty: I take the discussion about the so-called "causal theory of reference", developed by Putnam and Kripke in the early 1970ies, to have shown two things: A) We want to grant causal connections between tokens of some kinds of symbols and their reference a role in the determination of the meaning of the symbols – in particular, we want to do this in the case of natural kind terms, such as "gold", where the stuff they refer to, the element *gold*, plays a role in the determination of what counts as gold and what does not. This is what Putnam called the "contribution of the environment". (I say "we want to grant" because it is important to see that Putnam's and Kripke's discoveries are discoveries about our linguistic intuitions.) B) The causal relations between, for example, the tokens of "gold" and the element *gold* are immensely complex and it is extremely hard to figure out the particular causal relation that should connect the token to its referent. A given token stands in any number of causal relation and none of these by itself distinguishes itself as the right one ("gold" does not refer to jewelers shops or to chemistry textbooks or to metal or to undiscovered fake gold). What we need is a notion of "explanatory cause", the cause that is relevant for our explanatory intentions.[79]

The Putnam/Kripke story shows that the causal relation of a *linguistic* symbol to its referent must involve the intention of speakers to refer to a

79 What is relevant here is not so much semantic externalism (that has lead to externalism about mental states) but Putnam's critique of his own earlier causal theories of reference. This critique shows that a successful story of the causal relation between my tokens of "gold" and gold has to involve my desire to refer to that particular metal with that particular word. Putnam has tried to show this in his model-theoretic argument (1981, 34 etc.) and in the point that we need to single out what we mean by "cause", given that any event has several causes - whereas we need the one "explanatory" cause (Putnam 1982, 1985). This is supported by Wittgensteinian arguments to the effect that deixis is necessarily ambiguous (sometimes called the "disjunction problem"). When Kripke pointed at the cat (and Quine's native pointed at the rabbit), were they pointing at a cat, a feline, an animal, a flea, a colour, or a symbol? When Putnam pointed at water, how much H2O did we need in the sample for reference to be successful?

specific object or kind: otherwise it is underdetermined due to the multiplicity of causal chains.
In earlier papers (Müller 2004, Raftopoulos and Müller 2006, forthcoming 2006), we investigated the psychological evidence that there are simple input mechanisms in vision that result in an "object file" (Kahneman & Treisman & Gibbs 1992), which could be used for reference grounding without presupposing higher cognitive mechanisms. These mechanisms are bottom up and cognitively encapsulated, so they could presumably be present (if they are computational) in a pre-cognitive computational system. However, even such simple mechanisms require that the system has *attention* directed at an object in sensation in order to differentiate it from the background and from other objects (cf. Raftopoulos 2006, 55ff). They are thus not purely syntactic.
To put this in terms of meaning acquisition: How can a system acquire meaningful symbols without making use of cognition? Could there be a theory of language acquisition (or machine learning) that assumes a language can be learned by a system that has no cognitive processes? I propose that to develop such a theory is more than just a challenge: it cannot be done.[80]

Relation to Searle's "Chinese Room Argument"

Let us illustrate the same point in the terms used in Searle's "Chinese room argument". Searle's central notion is "understanding" (of Chinese and of stories) and he claims, 1) that the symbol manipulator in the Chinese room should not be said to *understand* Chinese by virtue of his handling the symbols correctly and thus producing correct output, also that he has no chance of learning Chinese [both of this everybody agrees with], 2) that the whole system containing the Chinese room, with manipulation manuals and all, cannot be said to understand Chinese [the "systems reply"], not even if "sensory organs" (cameras, microphones, etc.) are added [the "robot reply"], since these supply "just more Chinese". He sometimes expresses this as saying that the system has syntax but no semantics for its symbols: that symbols in a system cannot acquire meaning due to mere symbol manipulation.

---

80 Fodor's recent battle against behaviorist accounts of concept possession fires back on his Cartesian theory, when he insists on the problem that knowing how to apply "trilateral" is necessarily also knowing how to apply "triangular", even in counterfactual cases (Fodor 2004, 39), since whatever thing typically causes an instance of "triangular" also causes an instance of "trilateral". This is worse than Quine's undetached rabbit parts and, of course, than the rabbit fly as reference for "gavagai".

As several people have pointed out, 2) does not follow from 1). The upshot of the argument is, in my view, that Searle sets the task to explain *how* a system can understand Chinese given that the central symbol manipulator does not. After the Chinese Room Argument the belief that a symbol manipulating system can "understand" is in doubt and would require positive support.

Searle's claim is that he cannot learn Chinese by manipulating the symbols in his room, even if he tries hard – and then he expands this point to the whole system. But he already grants too much: Searle in the room *understands* the symbols in the instructions for manipulation, *wants* to learn Chinese, *knows* that Chinese is a language, that some of its symbols refer and which world they refer to. None of these is given in an actual purely syntactic computational system. Given that there is literally *no* understanding, desire and knowledge in the actual Chinese Room of a syntactic system (there are no intentional states), there is even less reason to believe that there is in the whole system.

The argument presented above thus goes some way towards closing the gap in Searle's argument by explaining why symbol manipulation, even under causal interaction with the environment, cannot produce intention. The system will not acquire meaningful symbols because it lacks everything necessary, specifically it has no *desire* to do so (it has no desires directed at anything). The situation is thus worse than in Searle's "Chinese Room", where Searle tries to show that an intelligent agent operating a purely syntactical system *cannot* acquire meaning. We only need to claim that a purely syntactical system itself *will not* acquire meaning - even if it could.

On a cautionary note, just like Searle, we do claim to have found any bounds as to what can be done with purely syntactic computing. Clearly, advanced AI systems (and perhaps "lower" animals) have achieved impressive feats without the "meaningful symbols" we have been asking for and which humans surely possess.

This look into the Chinese room might leave a paradoxical air, one might wonder what that magical bit is which allows humans and other animals what computers cannot have. My suggestion here is that this bit has to be something that is not computational – and I think *desire* is a good candidate.

Taking Stock: Some Conclusions

What we have seen so far is that:

1) A language of thought computational system presupposes meaning

2) A purely syntactical computational system is possible
3) A purely syntactical computational system could only acquire meaning if that process does not involve any mental states with intention (e.g. desires, beliefs, attention, …)

What we have not seen is whether there is another version of computationalism that could save the day. Perhaps there is computation without symbols or there is information processing in ways other than computing? Let us take a brief look at the options.

Vacuous Computationalism

Searle has repeatedly said that whether a system is a computer or not depends on its interpretation by some observer, a syntactic property is an observer-relative notion (cf. Preston 2002, 42-44). This is why he comes to the prima facie surprising conclusion that "The brain is a computer, in the sense that it instantiates computer programs," because "everything is a digital computer at some level of description" (Searle 2002, 224).

Whether this view is true or not (I tend to think it is not; see Piccinini, forthcoming2, for a detailed discussion), as Searle knows, this makes computationalism vacuous. Clearly, computationalism cannot be the claim that, if an observer likes to see it that way, the brain is a computer, and so is a train, a tree or a bumblebee.

Non-Symbolic Computing and "Information Processing"

There are cognitive scientists that use the word "computational" in a much weaker sense than the one defined above - in fact, the plethora of definitions is depressing: I counted 9 different ones in a recent exchange between Pinker and Fodor (2005).[81]

---

81 Fodor and Pinker in their 2005 exchange alone use the following, most of which are obviously either too narrow or too broad:

1) Literally being a Turing machine with tape and all (attributed to Fodor by Pinker 2005, 6). Falsely attributed and failing to mention that the relevant notion is that of the "universal" Turing machine.

2) "Cognitive architecture is Classical Turing architecture" (Pinker 2005, 6). If "architecture" is taken sufficiently abstractly this is different from 1). But what is that "architecture"? Perhaps being able to "compute any partial recursive function, any grammar composed of rewrite rules, and, it is commonly thought, anything that can be computed by any other physically realizable machine that works on discrete symbols and that arrives at an answer in a finite number of steps" (Pinker 2005, 6) on Turing). But this is a description of abilities, not of structure.

One prominent idea is that computing is somehow "information processing". But information processing could take many forms, some of which are not computational. There are many systems that could be used to compute but should not be called a computer. Dynamical systems in the sense of van Gelder (1995) are one example. Another are analogue systems, such as slide rules, mechanic (non-digital) adding machines, scales, tubes, etc. So, even if computing is information processing, what distinguishes it from *other* forms of information processing - some of which may even produce the same results? Surely this must be the *mechanism* by which it achieves that processing: namely computation (i.e. performing algorithms).[82]

I would therefore make the terminological suggestion to distinguish between "computationalism" and "information processing" as paradigms for cognitive science.

---

3) Having "the architecture of a Turing machine or some other serial, discrete, local processor" (Pinker 2005, 22 - attributed to Fodor). False attribution, since in 2000, Fodor did not mention the possibility of other processors. Suggests that "architecture" means physical setup (tape and reader), after all – see problems in 2).

4) Being 'Turing-equivalent', in the sense of 'input-output equivalent' (Fodor 2000, 30, 33, 105n3). Surely too weak. Any information processing system is input-output equivalent to more than one Turing machine.

5) Being 'defined on syntactically structured mental representations that are much like sentences' (Fodor 2000, 4). "Defined on" and "much like sentences"? A definition of the language of thought? Not of computation, surely.

6) Being supervenient "on some syntactic fact or other" - "minimal CTM" (Fodor 2000, 29). Too minimal, as Fodor himself agrees.

7) Being "causally sensitive to, and only to, the syntax of the mental representations they are defined over" [not to meaning] AND being "sensitive only to the local syntactic properties of mental representations" (Fodor's upshot in 2005, 26) - delete "mental" above and note that none of this makes for a computational process.

8) "In this conception, a computational system is one in which knowledge and goals are represented as patterns in bits of matter ('representations'). The system is designed in such a way that one representation causes another to come into existence; and these changes mirror the laws of some normatively valid system like logic, statistics, or laws of cause and effect in the world." (Pinker 2005, 2). Any representational systematic process is computational, then.

9) "... human cognition is like some kind of computer, presumably one that engages in parallel, analog computation as well as the discrete serial variety". Pinker 2005, 34 on Pinker - note the "like", "some kind" and "presumably", plus the circularity of using "computer"!

82 There are at least two notions of algorithm possible here, depending on whether the step-by-step process is one of symbol manipulation or not. (E.g. Harel 2000 introduces the notion of algorithm via a recipe for making chocolate mousse.)

Outlook: Analogue and Hybrid systems

The pages above present a reason to believe that the mind is not a computational system. However, there is still good reason to think that some parts of the human mind are computational, even if the problems explained show that it is not *only* that. Perhaps the picture that emerges is that of a hybrid and modular mind where some modules are computational but many are not. Some of the non-computational systems will be mathematically describable, perhaps in continuous mathematics, and can thus be simulated on digital computers to some degree of accuracy.

References

Chalmers, David J. (1993) "A Computational Foundation for the Study of Cognition" Online at http://consc.net/papers/computation.html

Churchland, Paul (2005) "Functionalism at Forty: A Critical Retrospective". *Journal of Philosophy* 102/1: 33-50.

Copeland, David (1993) *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

— (1997) "The Broad Conception of Computation". *American Behavioral Scientist* 40: 690-716.

— (2000) "Narrow Versus Wide Mechanism, Including a Re-Examination of Turing's Views on the Mind–Machine Issue". *Journal of Philosophy* 97/1: 5–32. (Reprinted in Scheutz 2002.)

— (2002) "Hypercomputation". *Minds and Machines* 12: 461–502.

Davis, Martin (2000) *The Universal Computer: The Road from Leibniz to Turing*. New York: W. W. Norton.

Fodor, Jerry (1981) "The Mind-Body Problem." *Scientific American* 244. Reprinted in John Heil (ed.) *Philosophy of Mind: A Guide and Anthology*. Oxford, Oxford University Press 2004, 168-182.

— (1994a) *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge, Mass.: MIT Press.

— (1994b) "Fodor, Jerry A.", in: Samuel Guttenplan (ed.): *A Companion to the Philosophy of Mind*, Oxford: Blackwell, 292-300.

— (1998) *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press.

— (2000) *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*, Cambridge, Mass.: MIT Press.

— (2004) "Having Concepts: a Brief Refutation of the Twentieth Century", with "Reply to Commentators", *Mind and Language* 19, 29-47, 99-112.

— (2005) "Reply to Steven Pinker 'So How Does The Mind Work?'" *Mind & Language* 20/1: 25-32.

— /LePore, Ernest (1991) "Why Meaning (Probably) Isn't Conceptual Role", in: Fodor/LePore *The Compositionality Papers*, Oxford. Oxford University Press 2002, 9-26.

Gärdenfors, Peter (2000) *Conceptual Spaces: The Geometry of Thought.* Cambridge, Mass.: MIT Press.

Harel, David (2000) *Computers Ltd.: What they Really Can't Do*. Oxford: Oxford University Press.

Harnad, Stephen (1990) "The Symbol Grounding Problem". *Physica D* 42 : 335-346.

Hauser, Larry (2002) "Nixin' Goes to China". In Preston & Bishop 2002, 123-143.

Haugeland, John (1985) *Artificial Intelligence: The Very Idea.* Cambridge, Mass.: MIT-Press.

— (2002) Syntax, Semantics, Physics. In Preston & Bishop 2002, 379-392.

Hofstadter, Douglas R. (1979) *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Basic Books.
Kahneman, D., Treisman, A. and Gibbs, B.J. (1992): “The Reviewing of the Object Files: Object-Specific Integration of Information.” *Cognitive Psychology* 24: 174-219.
Kim, Jaegwon (1996) *Philosophy of Mind*, Boulder, Col.: Westview Press.
Lycan, William G. (2003) “Philosophy of Mind”, in: Nicholas Bunnin/E. P. Tsui James (eds.): *The Blackwell Companion to Philosophy*, 2nd revised edition, Oxford: Blackwell, 173-202.
Müller, Vincent C. (2004) “There Must Be Encapsulated Nonconceptual Content in Vision”, in Athanasios Raftopoulos (ed.), *Cognitive Penetrability of Perception: Attention, Action, Attention and Bottom-up Constraints*, Huntington, NY: Nova Science, 181-194.
Newell, A. and Simon, H.A. (1981) “Computer Science as Empirical Inquiry: Symbols and Search.” *Communications of the Association for Computing Machinery* 19, 113-26.
Piccinini, Gualtiero (forthcoming1) “Computation without Representation”, forthcoming in *Philosophical Studies.*
— (forthcoming2) “Computational Modeling vs. Computational Explanation: Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?”, forthcoming in *The Australasian Journal of Philosophy.*
Pinker, Stephen (2005) “So How Does the Mind Work?” and “A Reply to Jerry Fodor on How the Mind Works.” *Mind & Language* 20/1: 1-24, 33-38.
Preston, John (2002) “Introduction”. In Preston & Bishop 2002, 1-50.
Preston, John & Bishop, Mark (Eds.) (2002) *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford: Oxford University Press.
Putnam, Hilary (1981) *Reason, Truth and History*. Cambridge: Cambridge University Press.
— (1982) “Why There Isn’t a Ready-Made World”. In *Realism and Reason: Philosophical Papers, Vol. 3* (pp. 205-228). Cambridge: Cambridge University Press.
— (1985) “Reflexive Reflections”. In *Words and Life.* Cambridge, Mass.: Harvard University Press 1994, 416-427.
Raftopoulos, Athanassios (2006) “Defending Realism on the Proper Ground”. *Philosophical Psychology* 19/1: 47–77.
Raftopoulos, Athanasios, and Vincent C. Müller (2006) "The Phenomenal Content of Experience", *Mind and Language* 21 (2): 187-219.
— (forthcoming 2006) “Deictic Codes, Object Files and Demonstrative Reference”, *Philosophy and Phenomenological Research*.
Rey, Georges (2002) “Searle’s Misunderstandings of Functionalism and Strong AI.” In Preston & Bishop 2002, 201-225.
Schneider, Uwe & Werner, Dieter (2001) *Taschenbuch der Informatik*. 4th ed. Leipzig: Fachbuchverlag Leipzig.
Scheutz, Matthias (2002a) (ed.) *Computationalism: New Directions.* Cambridge, Mass.: MIT Press.
— (2002b) “Computationalism: The Next Generation”, in Scheutz 2002a.
Searle, John (1980) “Minds, Brains and Programs”, in *Behavioral and Brain Sciences* 3: 417-457.
— (1984) “Intentionality and Its Place in Nature”, in: Searle *Consciousness and Language*, Cambridge: Cambridge University Press 2002, 77-89.
— (2002) *Consciousness and Language*. Cambridge: Cambridge University Press.
Smolensky, Paul (1999) “On the Proper Treatment of Connectionism”. *Behavioral and Brain Sciences* 11: 1-23.
— (1994) “Computational Models of Mind”, in: Samuel Guttenplan (ed.): *A Companion to the Philosophy of Mind*, Oxford: Blackwell, 176-185.
Van Gelder, Timothy (1995) “What Might Cognition Be if Not Computation?” *The Journal of Philosophy* 91/7: 345-381.
Wakefield, Jerome C. (2003) “The Chinese Room Argument Reconsidered: Essentialism, Indeterminacy, and Strong AI”, *Minds and Machines* 13: 285-319.

DOING WELL IN THE IMITATION GAME:

CASES PRO AND CONTRA ARTIFICIAL INTELLIGENCE RECONSIDERED

Hajo Greif

ICT Research Unit

Inter-University Research Centre for Technology, Work and Culture (IFZ),

Schloegelgasse 2, 8010 Graz, Austria

+43-316-813909-29

greif@ifz.tugraz.at

Asking the Right Question

A certain sense of disillusionment with the bold claims and bright prospects of early artificial intelligence (AI) research has become the general mood among its younger practitioners. Or so it seems. Machine consciousness and human-like intelligence are not an issue anymore. The design and implementation into digital machines of much more modest tasks have become the daily business.

Not so for philosophers. Most of their debates still seem to keep revolving around the question proposed in the opening sentence of Turing's classic paper "Computing Machinery and Intelligence" (1950), that document foundational to both AI research and the philosophical

debates about the subject. That question was: “Can machines think?” However, what struck me on reading and re-reading Turing’s paper was that he himself, although proposing that question, carefully refrained from really *asking* it, choosing a different question instead.

Often seemingly ignoring Turing’s own reservations, two possible interpretations of the original question, a stronger and a weaker one, have been discussed at length by philosophers of AI.

Granted that the doctrine of materialism holds, on the stronger interpretation creating artificial intelligence amounts to *creating a machine with cognitive features human-like in all relevant respects*, albeit in a different medium.[83] The condition of human-likeness in all relevant respects could be pushed to indefinite extremes by the sceptic who argues that the activity of thinking is too complex, and too deeply embedded in the human condition to be removed from its context and reverse engineered in stand-alone fashion.

In various incarnations, and with different degrees of sophistication, this sceptical bottom line is continued through many of the philosophical critiques of AI. This kind of essentialist objection could only be fended off pragmatically by selecting some relevant respects in which machines shall be like humans, on the background of significantly more limited explanatory goals. If the demand for identity is too strong, AI settles for the task of *creating machines whose internal operations represent the key features of human thought processes*. That representation would provide an explanatory model of the structure of, and the processes within, the human mind, while leaving open whether this amounts to a re-creation of it, that is, a thinking mind in itself.

The way in which an AI system operates, and the means it uses, on this view is an experimental design for testing theories about the human mind. Its component parts, in their specific arrangement, with their specific effects, are empirical models of theoretical entities within that theory, allowing for predictions as to what cognitive psychologists and neuroscientists will find out about the inner workings of the human

83 The claim that AI were acceptable if it were about creating artificial human beings is to be found in Searle (1980), p.422—although here, of course, it is meant as a criticism of AI.

nervous system.[84] For comparison, one might consider the double helix model of the DNA which, although DNA structures were not directly observable to Watson, Crick and Franklin, and, of course, although the model was constructed of different materials on a different scale, and was never meant to replicate like DNA, allowed for predictions about how DNA sequences would look like on future observations, and what effects future experiments would show. Of course, one could also think of models that indeed represent aspects of the behaviours of the entity modelled, but this is not an essential criterion for this kind of design. How much of the mind's structure and operations could be represented in this fashion, makes the difference between the classical "strong" and "weak" paradigms of AI.[85]

In any case, a preceding question that needs to be answered, in order to formulate a theory in the first place, is this: What are the salient features of human thought processes that should be represented in a computer, and how should they be so? There are various suggestions:

(R 1) A computer program could be designed so as to provide the basis of a model of the human mode of operating on symbols as mental tokens—however it may be realised in terms of computational architecture.[86]

(R 2) A computational architecture could be designed to map onto the interrelations between the functional elements within the human nervous system, and their dynamics—however inferences, feelings, memories etc. may find their place within that network.[87]

---

84 In formulating this point so tortuously I am trying to keep up with the important, but often forgotten distinction between programs, models, and theories in the concept of computer modelling, as discussed in Moor (1978), pp.219-222. The author argues that a computer program, taken by itself, although being based on a theory and providing a model of the explanandum, is neither a theory nor a model of what shall be computer-modelled. Only the entire system, as it operates, may serve as a model, in accordance with an independently formulated theory.

85 This is an interpretation of Searle's classic distinction, in (1980), p.417: Strong AI is the programme of making AI systems the explanation of human cognitive processes in their entirety. Being such an explanation, Searle argues, in the view of Strong AI defenders amounts to being a thinking mind itself—which indeed would be a premature conclusion. Weak AI, on the contrary, is the programme of generating experimental structures on which some hypotheses about the human mind could be tested, but it might not behave like the human mind at all.

86 This is the classic idea of "computer simulations of human thought processes" as defined in Feigenbaum/Feldman (1963).

87 This is the basic idea of connectionist AI, as described in Rumelhart/McLelland (1986).

(R 3) The most demanding suggestion is to combine representations of both the program and the architecture level in a way that captures the properties of each level (R 1 and 2) as well as the interrelations between them.[88]

In any of these varieties, AI amounts to the programme of exploring and explaining the inner structure of the human mind as precisely as possible, by identifying the properties within it that are *necessary* for it to think. By implication, this programme, if successful, would provide an explanation of what it *means* to think. However, even in its more modest varieties, this programme remains vulnerable to the objection that it cannot explain the structure and the functions of the human mind at all if it selectively removes them from their organic and/or life-worldly context. What it is and what it means to think could not be understood without looking at what that thinking does, and what is done with that thinking, in the natural and social environment in which it takes place.

Reasons for Playing the Imitation Game

My suggestion here is that the approach to AI I just outlined was not Turing's programme, and that the alternative he hinted at is apt to contain the latter objection. Whether or not it was because he saw those difficulties coming, Turing expressly refrained from a literal reading of the question of thinking machines. With the common sense meanings of the words "machine" and "think" current at his time, he observed, the question would turn out to be unanswerable simply because it is meaningless.[89] However, unlike many philosophers, he also refrained from embarking on a conceptual analysis of the notion of thinking that would give his question a viable meaning, and the subsequent research programme a clearly defined target. What Turing did instead was to propose an experimental design to test tentative and, in fact, rather modest theoretical definitions of thinking that could be implemented in certain machines. His aim was to validate the functional layout and the capabilities of the machines as limited analogues to human thought

88 Proposals towards such a 'hybrid' approach are to be found in Hinton (1991).

89 See Turing (1950), p.442.

processes in the first place, rather than to explain what thinking really is.[90] However, some interesting suggestions towards the latter emerge on a secondary level.

Turing's experimental design is reflected in his interesting rephrasing of the original thinking-machines question that has the virtue of being far less demanding than the original, but answerable instead: "Are there imaginable digital machines that would do well in the imitation game?"[91]—where the imitation game consists in a form of partially blindfolded communication which is set up so as not to allow the human partner to identify the true, human or machine nature of his or her counterpart.

There are two interrelated aspects to the design of that thought experiment:

(i) developing a digital machine (i.e. computer or robot) that reliably operates on (human-generated) behavioural inputs in a way that produces an output in some respects indistinguishable from human behaviour itself;
(ii) devising channels of human-machine interaction that are narrow and standardised enough to secure indistinguishability of the machine behaviours from human behaviours and expressions.

The more refined the computing machinery and its programming, the less need there would be for concealing the digital nature of the counterpart in the imitation game. Neither identity with, nor representation of the structure of human cognitive processes is demanded. Instead, Turing's idea was a different one: It refers to the possibility of *simulating intelligent human behaviour in some relevant respects*. Those respects are selected in the set-up of the imitation game, with the presumed result of partial equivalence of behaviours under experimental conditions.

While the design of the thought experiment may suggest otherwise at first, the imitation game does not amount to machines simply *mimicking* human behaviour. Such mimicry, as in the case of ELIZA, would be the case if the machines in question produced a certain set of anticipated

90 This is where Gunderson's otherwise well-reasoned critique of Turing in (1964) misapprehends the latter's explanatory purpose: It is only some aspects of thinking that shall be simulated in the imitation game.

91 Turing (1950), p.442.

effects that are tailor-made to look like human behaviours if communicated through the appropriate channels. Instead, the imitation game will have to remain operational even in unanticipated cases that were not hard-coded in advance. Thus, some capability of learning and developing the scope and depth of simulated behaviours is required—which Turing fully acknowledges.[92]

In order to be properly assessed, the “imitation game” question requires an answer to another, preceding question: Which aspects of *human* behaviour are the cues to its intelligence, thereby allowing for a decision on what shall be simulated, and how, given the machine’s limitations as an imitator? There are different suggestions:

(S 1) Uttering apparently meaningful, true, and contextually adequate subject-predicate sentences, or at least equally adequate sequences of formal symbols, seem to be a good candidate for the behavioural cue to intelligence that needs to be included. (That was Turing’s own suggestion.)

(S 2) On the other hand, informal cues like prosody, facial expression, or gestures may be equally, or even more important under certain circumstances, while being all but impossible to be simulated by a computer, and still difficult to implement in a robot.

(S 3) But perhaps the best candidate to begin with is the basic, sub-linguistic competence of behavioural adaptability in novel situations, on which the other capacities (S 1 and 2) build up.

In any of these cases, like in simulations in general, the aim is to recreate a selection of observable features of some process or behaviour by (mostly technological) means different from those present in that process or behaviour.

The purpose of a simulation may be, on one approach, that it can be interacted with in much the same way as one would interact with the original, simulated system. Following Ringle (1979, p.7 f), these might be called “demonstrative simulations”. Flight simulations may provide a good example: sophisticated flight simulators may give a vivid impression of flying even to an experienced pilot, and pose challenging, flight-typical problems for her to solve, yet their system architecture and their physical

92 See the concluding observations about machine learning in Turing (1950), pp.454–460.

features have little in common with an aeroplane. (And, of course, it would be silly to ask whether they can fly.)

Alternatively, on another approach, a number of selected behaviours of a system may be recreated in order to identify their effects and, if applicable, their function. These may be called “investigative simulations”. Outright similarity on the observational level is not required here, but the careful mapping of the effects and/or functions of the original. Wind channel simulations of aerodynamic designs or computer simulations of energy flows in a building may be listed as examples. From this kind of simulations, one primarily wishes to learn about the regularities in the patterns of behaviour of the simulated systems. (And, of course, it would be silly to assume that one could put the simulation in place of the original.)

In either case, the criterion for a successful simulation is not that it represents the internal structure of the original, but that its input-output relations are equivalent to those in the original.[93] On this view, an AI system is an experimental design for testing theories about the human mind inasmuch as, and only inasmuch as it seeks to identify *sufficient* conditions for intelligence, not the necessary ones, that is, it asks what kinds of systems may deliver identical, or in some respect comparable, behavioural effects on equivalent inputs. In the imitation game, equivalence of input-output relations amounts to the correspondence of the behavioural simulations with with natural human behaviours, so as to create patterns of, on some level, seemingly rather natural interactions.

---

93 One classic definition of AI is “to construct computer programs which exhibit behaviour that we call ‘intelligent behaviour’ when we observe it in human beings.” (Feigenbaum/Feldman, 1963, p.3, emphasis in original). Given this definition, one would be justified in calling a machine “intelligent” if it repeatedly and reliably exhibits behaviour in this very fashion. Nothing is said here about how this behaviour is achieved. It seems that the focus and even the definition of AI research has much changed after early attempts like Turing’s own and this one.

Thinking Functions

The nature and advantages of the simulation approach are perhaps best understood if we first reconsider two of the most persistent critiques of AI, to which I attached the name tags of their main proponents:

"Searle": Thinking machines are impossible because the intentionality of their candidate thoughts, qua being representations of thoughts at best, inevitably is only of derived nature, whereas natural minds, qua natural, possess original intentionality, by virtue of their thoughts being about the world.[94]

"Dreyfus": Thinking machines are impossible because their candidate thoughts, qua being abstractly modelled, do not emerge from interaction with the machine's environment, whereas the natural minds' capacities exist precisely because they are situated in an environment with which they interact.[95]

While the authors of these critiques, when it comes to assessing the privileged nature of natural minds, resort to obscure "causal powers" or to Heidegger respectively, there is an etiological, evolutionary argument that may help to both elucidate and connect those critiques, while at the same time allowing for the possibility of artificial intelligence, albeit of a different design.

In a nutshell, the story is this:[96] To be a natural thinking being is one of several possible ways of adaptively acting within an environment. However, an organism's environment is not merely its surroundings, but the specific set of conditions relevant to its survival and reproduction. These may be different for different species, for example acidity of the medium and sufficient daylight, water and warmth for some plant, vs. the presence of the former plant and the absence of predators for some herbivorous animal inhabiting the same place. Therefore, the organism's

---

94 This critique is paradigmatically formulated in Searle (1980). At some point in his essay (p.424), he stops surprisingly short of an evolutionary explanation of intentionality as a biological phenomenon of the kind I am proposing here.

95 This critique is proposed in Dreyfus (1992). It is interesting to see that it is mobilised in defences of the behaviour-based AI approach, but now in order to identify a new set of criteria for the possibility and design of intelligent machines, like in Hallam/Malcolm (1994); see also Dennett (1994), p.142 f.

96 I am relying here on naturalistic theories of the mind as those proposed in Millikan (1984).

relations to, and interaction with that specific environment, based on the evolutionary history that has shaped that organism and fitted it into its current environment, define its organic and behavioural traits and their functions. Mostly, this defining happens by way of variation and natural selection, as differential reproduction under some specific set of environmental conditions. By implication, these processes also define whether, to which degree, and by which means an organism represents to itself its own relation to that environment. In some organisms—presumably those for whom a complex set of environmental conditions is relevant—, behavioural traits to some degree are shaped by learning histories, that is, by internal selection of behavioural patterns according to memorised results of earlier behaviours and/or projected goals. These activities require some form of internal representation within the organism.

In this sense, human intelligence is a natural capacity, just like many others, with, qua having been positively selected, its own adaptive functions. It resides upon, and is historically derived from, older, more basic biological functions, all or most of which contributed to the ancestors' survival. Those functions, by virtue of having been reproductively established, are the organism's proper functions.[97] For artificial systems, "reproductively established" means: being part of successful, as it were, ancestor designs that are incorporated into more recent ones.

In the cognitivist AI programme, interactions with the environment, both past and present, are left out in favour of the quest for accurate representations of some specific operations within the organism, which are themselves thought of as, first and foremost, representations. These representations are conceived of in an abstract and formal fashion, not as responses to the needs of the entire organism. Yet these operations and their functions would not exist without the challenges the environment poses to the organism.

---

97 For an exposition of the latter, see Millikan (1984), pt. I. Note that Millikan explicitly states (p.1 f) that this concept is meant to apply both to natural and to artificial systems (although she only considers technical artefacts in general): A proper function is the effect of a system produced frequently and reliably enough so as to account for its continued reproduction. On this definition, the intended functions of artefacts may differ from their proper functions.

In fact, for robots that were designed along traditional cognitivist AI principles, it proved rather difficult to achieve successful interaction with a natural environment in which they were placed. Even in the case of seemingly simple navigation tasks, the problems for them were, firstly, to autonomously identify and re-identify objects, secondly to recognise their behaviourally relevant properties, and, in consequence, thirdly, to create cognitive maps and successfully navigate by them. Systems that make hardly any use of inner representations, but were designed to interact appropriately proved more adaptable and versatile—precisely because they were not designed to represent, but to interact.[98]

One may try to capture the essence of behaviour-based AI by saying that it abandons the notion of representation in a twofold way: In the first place, it does not think of the project of AI anymore as the endeavour of representing the inner workings of the mind. But secondly, it does not think of representation within the systems in question as the explanatory key to intelligence altogether. Instead, it seeks is to capture the appropriate relations between system and environment that require behavioural adaptations from the former. Such relations begin with (only seemingly simple) tasks like navigation, orientation, and avoidance, not with proving theorems or understanding language.

To capture those relations, on this view, any artificial intelligence design will have to achieve *functional equivalence* with the mental operations of a natural system, which it could only do if the artificial system’s own operations relate to its own environment in roughly the same manner as the natural system's mental operations relate to the natural system's environment, and make a comparable contribution to its continued existence. In which particular way this is achieved is of secondary importance at most. What really is important instead is to identify the purpose for which a certain trait is, or has been, selected. This is what has to be equivalent between the natural and the artificial system. However, the same purpose can be fulfilled by a variety of different means, without principled limitations. Functions are underdetermined by definition.

98 This is the basic idea behind behaviour-based AI, as outlined in Maes (1993); Brooks (1991); Brooks (2002).

Thus, the AI system's traits, mental or other, may or may not be phenotypically identical with those of the natural system. What counts is that they make an equivalent contribution to sustaining the system, in spite of different, natural vs. engineering histories, and in spite of probably different internal structure. Consider the function of vision in different species, which is realised in a variety of eye designs made of different tissues, serving as models for camera-based vision in robots.

Accordingly, in order to simulate human intelligence, it is precisely its functions, and not its structure or contents, that need to be successfully identified and re-created in a different medium. To identify those functions is to repeatedly observe, in as many cases as possible, and including a perspective on early human history, the relation between the behaviour of human beings and their environment, including other human beings, and to make a qualified judgement as to whether, when and in which ways behaviours generally succeed, and when they fail.

Routes to Intelligent Behaviour

If, on these grounds, AI is possible in terms of functional equivalence, that is, behavioural traits being selected for comparable effects, the route to that functional equivalence still may lead to—at least—two widely different designs adapted to two different purposes, which seem to map onto the distinction between investigative and demonstrative simulations outlined above:

(AL) AI systems may be designed, at the most basic level, as in (S 3) above, *to interact with their environment in the same fashion as a natural organism.* Their task is to autonomously and adaptively navigate in a natural, changeable environment in real-time—avoiding obstacles, identifying and following moving objects, finding humans or other machines to interact with. Outright similarity in appearance of the system or its behaviours to real-world organisms, let alone humans, is not required for such systems to pass as functionally equivalent, since the same kind of functions may be realised in apparently different designs. What is required is the mapping of the effects and functions of the simulating on the simulated system, however simplified those effects and functions may be in the simulation. Nor are complex representations

essential for this kind of AI system. This “animat”, or “artificial life” approach operates in a bottom-up fashion, which begins with the most basic behaviours, on which the higher faculties of the mind and their functions ultimately rest.[99]

(HR) AI systems may be designed, as in (S 2) above, *to interact with human beings in the same fashion as these would interact with each other.* The system’s verbal and facial expressions, prosody and looks are designed to evoke reactions in the people they encounter that would be typical for encounters among people. In turn, the system’s perceptual and motor facilities are adjusted to producing appropriate reactions to human verbal and facial expressions, prosody and looks in much the same way as would come from another human being. Experimental results in humanoid robotics indicate that machine interaction with human subjects to some extent quite neatly matches interactions between humans. Here, similarity of appearance and behaviours, and the right selection thereof, is a design criterion, because the naturalness of direct interactions between humans and machine is relevant—while autonomous navigation is not an issue at all, and while the computational architecture behind the machine is a purpose-built, heterogeneous computer network utterly dissimilar to the human perceptual, cognitive and motor apparatus.[100]

In either case, the machine’s successful interaction with its environment is the criterion for an operational AI design. That environment may be a space in which fixed and moving obstacles, and sources of light and energy are the only relevant conditions to interact with, and in regard to which the only means of interacting are moving either towards or away from them. That environment may also be, on the

99 At the MIT AI lab, robots were developed that were able to autonomously navigate in real-world environments, and to explore them, without even having any central processing unit, and therefore, arguably, having no representations of the environment they mastered. Sensors were directly linked to simple actuators, each unit interacting with the others only along a very simple algorithm, within a simple architecture in which modes of perception and behaviour are layered instead of centrally governed. That robot design, in its classic incarnation named “Genghis”, is described in Brooks (2002), earlier incarnations in Brooks (1991).

100 Again at the MIT AI lab, two types of humanoid robots were developed: one (“Kismet”) consisted in a desk-mounted head with some of the key features of a human face, each of them functional and movable: mouth, ears, eyes, and eyebrows, all designed to make Kismet interact with people by facial gestures, looks and prosody; another, earlier robot named (“Cog”) that may be an interesting case between the two paradigms outlined here, was equipped with a head with eyes and ears as well as with arms and hands, thus being able to grasp and look at, and eventually learn about, things he encountered. Both types are described in Brooks (2002); for more details on the Cog project, and philosophical observations on it, see also Dennett (1994).

other hand, the gestures, voices, looks and facial expressions of human beings, and nothing else, the interactions being restricted to precisely these forms of expression. The AI systems described above are successful players of the imitation game inasmuch as an appropriate set of relations with certain aspects of their surroundings was established for them, which make up the specific environments for the machines to function, and to which to adapt their behaviours.

However, a certain air of disappointment may remain: While humanoid robots do not anymore require Turing's restrictive channelling of interactions necessary for making them look natural in the imitation game, the functional equivalence, which I declared to be the issue here, remains questionable, since the behavioural traits and their functions do not really build up on historically older functions in ancestor designs towards a naturally 'viable' system, but are tailor-made, in top-down fashion, to achieve a certain effect. On the other hand, where functional equivalence is achieved in animats, it is so only on a rather rudimentary level—sometimes to the point of making the applicability of the label "artificial *intelligence*" questionable.

In the meantime however, the two approaches in conjunction may serve as blueprints for a truly valid test for the possibility of artificial intelligence: If animats may be developed that, strictly relying on a bottom-up design process, acquire cognitive functions equivalent to those of higher animals, and if those functions, with respect to interacting with the environment, are expressed in ways in which we recognise, and could relate to, at least a faint and partial simile of human patterns of interaction, the infamous explanatory gap in any programme of artificial intelligence will be ultimately closed. Just like people recognise each other as intelligent beings by their observable behaviours, similar behaviours in machines, based on the development of equivalent functions, may be cues to their intelligence.

But even if one remains sceptical—perhaps rightfully so, for the enormous practical requirements of such a project—, what the imitation game approach can still achieve is helping to explain the function of certain human behavioural traits in experimental fashion: What is really necessary in *human-human* interaction for recognising each other as intelligent beings? The selection of certain channels and the masking of

others may provide a good test for the observational cues people use, and therefore, in its own right, make a contribution to the cognitive psychology that Good Old-Fashioned AI was so desperately yearning for.

References

Brooks, Rodney: “Intelligence without Representation”. Artificial Intelligence, 47, 1991, 139–159.

– Flesh and Machines. New York: Pantheon, 2002.

Dennett, Daniel: “The Practical Requirements for Making a Conscious Robot”. Philosophical Transactions of the Royal Society of London, A, 349, 1994, 133–146.

Dreyfus, Hubert L.: What Computers Still Can’t Do. Cambridge/London: MIT Press, 1992.

Feigenbaum, Edward A./Feldman, Julian, editors: Computers and Thought. New York: McGraw Hill, 1963.

Gunderson, Keith: “The Imitation Game”. Mind, 73, 1964, 234–245.

Hallam, J. C. T./Malcolm, C. A.: “Behaviour: Perception, Action and Intelligence – The View from Situated Robotics”. Philosophical Transactions of the Royal Society of London, A, 349, 1994, 29–42.

Hinton, Geoffrey, editor: Connectionist Symbol Processing. Cambridge/London: MIT Press, 1991.

Maes, Pattie: “Behavior-Based Artificial Intelligence”. In: From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior. Cambridge/London: MIT Press, 1993, 2–10.

Millikan, Ruth Garrett: Language, Thought, and Other Biological Categories. Cambridge/London: MIT Press, 1984.

Moor, James H.: “Three Myths of Computer Science”. British Journal for the Philosophy of Science, 29, 1978, 213–222.

Ringle, Martin, editor: Philosophical Perspectives in Artificial Intelligence. Atlantic Highlands: Humanities Press, 1979.

Rumelhart, David/McLelland, J., editors: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge/London: MIT Press, 1986, 2 vols.

Searle, John R.: “Minds, Brains, and Programs”. The Behavioral and Brain Sciences, 3, 1980, 417–457.

Turing, Alan M.: “Computing Machinery and Intelligence”. Mind, 59, 1950, 433–460.

Section 2

psychology values & thought

HUMAN ORIENTED COMPUTATION:

THE UNIFICATION OF WITTGENSTEIN'S PHILOSOPHIES AS A GUIDE TO THE RATIONAL USE OF IRRATIONALITY

TOM ADDIS[101], DAVE BILLINGE

School of Computing, University of Portsmouth, Buckingham Building, Burnaby Road, Portsmouth PO1 3AE, UK. +44 (0)23 92 846400, Tom.Addis@port.ac.uk

Abstract The artificial intelligence community has still not fully addressed Dreyfus' criticisms of its attempts to emulate human reasoning [Dreyfus 1992]. We suggest that this is in part because we have not addressed the gap between the normal use of *dynamic* and *indefinable* set boundaries by people and the use of *clear* and *definable* set boundaries by computing machines. We show how Wittgenstein's two philosophies offer clear guidance for changing the technical paradigm so as to establish the use of the irrational set as part of the design and analysis toolkit of computer science and thus to enable the move to human-like computing machines. Surprisingly, this change in perception of the problem domain and the kind of demands it puts on technical solutions also has a strong impact upon ethical issues outside the sphere of computer science.

Introduction

The ambition to emulate human thought and intelligence on an electronic computer has been with us since their invention. Many advances have been made but we, the authors, are puzzled as to why,

101 Also Visiting Research Fellow, Science Studies Centre, Department of Psychology, University of Bath, UK

after sixty years of effort, millions of man hours of work and technology that boasts of silicon machines that do 1000+ Gigaflops with a 1000+ Terabytes of storage, computer science has still not emulated many of the important functions of the human brain; a device that looks like a bowl of porridge and consists of only 15 Gigacells working at about 50 cycles per second. We seem to be getting something wrong and we suggest that there need to be new ways of looking at the current problems. We propose a different way of viewing such problems that is drawn from the insights of Ludwig Wittgenstein. These insights suggest a novel perspective on the world, which leads to a different view on how computer programs should be designed. This will direct computation along different lines with new principles such as 'a structure malleable program' that adapts to the changing requirements of the human condition.

We expand on our initial proposal, given elsewhere [Addis *et al* 2004, 2005, Billinge & Addis 2004, Stepney et al 2005, 2006], that communication between man and machine requires dual semantics and we suggest that the only way we might bridge this gap is by providing a means of feedback so that the machine, operating exclusively within classical logic, may emulate human non-classical logic. We characterise the difference between classical and non-classical logic through the distinction between rational and irrational sets [Addis *et al*. ibid]. In this context we will give the term 'irrational' a very specific definition.

In order to provide a clear and rigid framework for our analysis we took the view from the Church-Turing Thesis [Kleene 1967] that a program can be considered as equivalent to a formal language similar to predicate calculus where predicates can be considered as functions. We have related such a calculus to Wittgenstein's first major work, the Tractatus, and we have used the Tractatus and its relationship to the world as a model of the formal classical definition of the semantics of a computer program. As such we will be referring to the details of the Tractatus as a general theory of semantics and programming languages. It will be a theoretical framework and a justification for all such formal languages. It is this framework that we will use to support our case for the limitations of computer programs.

The limitations of computer programs in particular arise from the essential flaw Wittgenstein found in his initial work when he considered how people use natural language. He explored these flaws in a new Thesis described in his second work; the Philosophical Investigations. We will use this later work as a new and extended theory to propose an approach to the designing of programs that are used within a human context.

We will show that natural communication with people will only be possible if we can find a mechanism that bridges the gap between the above two philosophies. The gap is a consequence of firstly the flaw identified by Wittgenstein, and secondly of the fact that computers depend upon formal languages to operate. In one narrow sense this means that the computer program must become part of society. By becoming part of society, we mean there must be a mechanism that can continually adjust to society's changing view of the world. We come to this conclusion by examining the way people communicate and by considering the work on the philosophy of language by Wittgenstein. We will highlight in particular the way in which Wittgenstein's two philosophies offer a developmental continuum that informs these efforts.

Finally, we note some important ethical principles that follow from this change of view and that impinge on the way we should govern our own society.

A Philosophical Paradigm and Computing

The computer and its program normally work within a fixed ontology, the rational paradigm, where all objects can be explicitly and exactly defined following the classical Three Laws of Thought: an entity is what it is and cannot change; an entity cannot exist both inside and outside the set boundary; and no entity can exist in neither the set nor outside the set boundary. This rigid requirement for computers and their programs is not required by normal human communication. As pointed out by Brian Cantwell Smith [Smith 1982] when comparing developments in programming and natural linguistics:

"…. computational semantics and linguistic semantics appear almost orthogonal in *concern*, even though they are of course similar in *style* ….."

He observes:

"It is striking, however, to observe two facts: First, computational semantics is being pushed (by people and by need) more and more towards declarative or referential issues. Second, natural language semantics, particularly in computational-based studies, is focused more and more on pragmatic questions of use and psychological import."

The focal problem for referential semantics is that it attempts to provide a *human independent mechanism* to assign meaning to a language. If this mechanism uses everyday objects or entities as referents, such as is done for Object Orientated Programming (OOP), then some well known paradoxes occur such as:

- The same physical object can adopt two distinct meanings, as in the morning and evening stars paradox.
- The destruction of a referent object causes a set of sentences that refer to that object, either directly of indirectly, suddenly to lose all sense because the referential object that provides the underpinning meaning to this set is no longer there to support the sense of it.

These paradoxes led Wittgenstein to his *Tractatus* [Wittgenstein 1921]. We took this work as a paradigmatic description of the current state of computer science because it avoids these paradoxes without commitment to any design methods such as OOP. We can take this step because the Church-Turing Thesis shows that the Turing Machine (the classical computer) is equivalent to Lambda calculus and recursive functions. Lambda calculus and recursive functions together are the underlying principles of a functional programming language (e.g. ML, LISP). Such a functional language is embodied in Wittgenstein's *Tractatus*[102]. This early work encapsulated a formal and logical

102 David Gooding (University of Bath, private communication 2004) notes that "the Tractatus was modelled on Hertz's Principles of Mechanics. Hertz believed that his book would be a full and final statement of the principles of mechanics; Wittgenstein thought that Frege, Russell and Whitehead had done the same for mathematics and that he would do the same for language."

representational schema into a descriptive form that is based upon denotational (or referential) semantics.

In order to avoid the paradoxes in a programming language and still retain the essential idea that meaning of a symbol (a name T3.202) is derived from its link to an object, the referents (T3.203) have to have some logically very strange properties. Objects must be:

- *independent* in that they can freely combine to form "states of affairs" that can be described (T2.01, T2.0122, T2.0124, T2.0272). The objects form relationships with each other (a configuration). This configuration is a state of affairs or a fact. All the existing state of affairs is the 'real world' or reality (T2.04, T2.06).
- *atomic* in that there are no smaller constituents (T2.02, T2.021). They are the substance of the world. Material properties in the world can only be produced by the configuration of these objects (T2.0231) in the same way that chemicals can only be formed by the configuration of atoms.
- *in all possible worlds* (T2.022, T2.023) since, no matter how strange a world might be, it must have something which is shared by the world we know.
- *immaterial* (T2.0231, T2.0233) because material properties are formed by the configuration of objects and not the objects themselves.
- *indescribable* except by their behaviour (form) (T2.0121, T2.021, T2.0271). For example, mass and force in physics can only be described in terms of their interaction with each other. Thus in the equation 'Force = Mass * Acceleration' it is only acceleration that can be observed. Space, time and colour (as to being coloured) are also only describable in terms of situations. These are examples of objects.
- *self governed* in that they have their own internal rules of behaviour (T2.012, T2.0121, T2.0123, T2.01231, T2.0141, T2.03, T2.033) in the same way atoms have valencies.

These referents (objects) are intended to be more than just elements of description; they form the real world (T2.04, T2.06). From these

referents, the full force of logic, predicate and propositional calculus retain stability of meaning and sense. Such a stance results in the position that everything is potentially unambiguously describable (T3.25, T7).

The *Tractatus* provides an extensive model, as well as a semantic theory, of computer languages. For example, the argument in the Tractatus, is that names (in practice signs; the visible part of an expression or name) in propositions do not always refer to primitive objects but are themselves referencing propositions (T3.14, T3.31, T4.22, T4.221, T4.03, T5.135 and further discussed in P43-60 [Wittgenstein 1953]). Thus, a Unicorn because it is a mythical animal cannot be an object and have meaning. However, it can be considered as a proposition, part of which is an animal with a single horn that is naturally centred on the forehead of a horse. For a computer programming language such propositions as that for a Unicorn are defined procedures or functions. These, in turn, can be complexes that finally end up as compound statements of assembler and then machine code whose ultimate referent is the bit[103].

This interpretation was also acknowledged, in a different context, by Bertrand Russell (May 1922). He noted in his introduction to the Tractatus:

> "that every language has, as Mr Wittgenstein says, a structure concerning which, in the language, nothing can be said, but that there may be another language dealing with the structure of the first language, and having itself a new structure, and to this hierarchy of languages there may be no limit."

This, of course, reflects the current state of affairs in programming languages, as we have discussed, where a macro language transforms into a high level language that compiles into assembler code, which in turn produces machine code. Even the structure of programs can be perceived as layers of language constructs [Visscher 2005].

---

103 For example, in computer languages we have seven bits of the ASCII code identifying 1000001 as the character A and 1000010 as the character B etc. There are also special characters such as 'delete' 1111111 and 'start' 0000001.

The Tractatus, as a theory, implies that the computer bit is the mechanical equivalent of Wittgenstein's referent objects. A 'bit' is a concept that can only be embodied in a distinction and it must have all the strange properties of Wittgenstein's objects, viz: it is *independent* in that each bit can change its state independently of any other; it is *atomic* in that there is no smaller element from which it is structured; it *exists in all possible worlds* in that no world can be imagined that does not have at least one distinction at least once in time; it is *immaterial* in that a distinction can be formed from any detectable difference such as electrical potential, positions of a bead on a string or the level of water; it is *indescribable* in that a distinction is either recognised or not but is always available to be used; and it is *self governed* in that the relationship between examples naturally forms a dimensional relationship. A particular 'bit' (example) according to the Tractatus will be an argument place (T2.0131). Further, it is through the bit that the program links to the world and has meaning. It is this meaning that allows the program to have "sense" with respect to the computer. This formal semantics and the ability for programmers to create procedures and sub-routines (sub-propositions or expressions) is the primary characteristic of all high level and assembler programming languages[104].

The consequence of taking the Tractatus as a formal model and theory is that any set of names can be used in a program to represent a proposition, procedure or function. All that is necessary is that there is a formal definition that gives the name sense within the program in terms of the proposition it represents. Since a proposition can take on an infinite number of forms through the use of tautologies (T5.142, T6.1, T6.12-T6.1-T6.1203) and other formal equivalences, then there is an infinite but bounded set of possible organisations that can be adopted for a program. The meaning of an *essential program* is bounded by such a set. We mean by an *essential program* some kind of base or minimum program that can be written that does the task. It is equivalent in concept to the most general unifier used in theorem proving. However, the additional adopted structure is also represented, in the end, by bits on a computer.

104 The original high level programming language COBOL in its initial form did not provide for procedures and sub-routines except those that were pre-constructed in assembler as library routines.

This will appear as a program overhead that is used to support a chosen program organisation or structure and in this sense the program interpretation has changed with such a reorganisation.

Dual Semantics

One of the explanations we will repeat here for completeness [Addis et al ibid] is that computer languages have a *dual semantics* in that the program signs (e.g. the names/labels given to data items, procedures and sub-routines) at the highest level also have referents in the world (figure 1 – the Problem Domain). This is the analysis of the problem domain in terms of records (as in database and program structures), relations (as in normalised data structures) and objects (as in OOP). It is this analysis that identifies constructs in the world that are meant to be stable and unchanging (as per *Tractatus* referents) to which names can be given and meaning assigned.

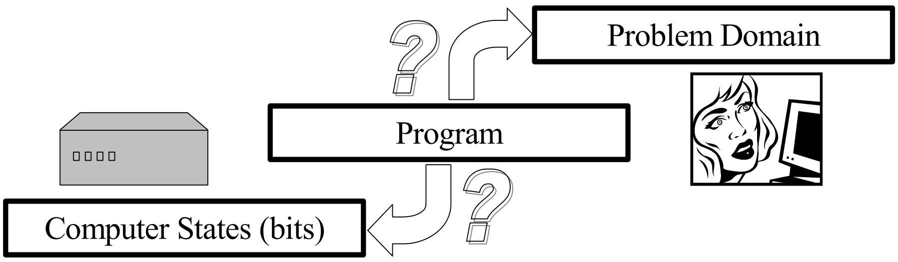


Figure 1. The problem of dual semantics

Now it is acceptable that propositions can represent material properties (T2.0231), relationships (T2.031), and any complex model of the world (T3.1-T3.32, T4.01, T4.021) *but a proposition can have one and only one complete analysis* (T3.25). Such an analysis is dependent upon only the essential features of the proposition (program) that link it to the referent objects (which is the bit in our case).

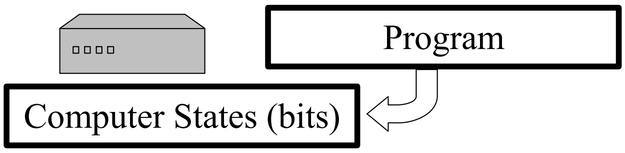


Figure 2 The only rational interpretation of a computer program

A computer program, as we have already seen, has such an analysis with respect to the computational engine (figure 2), so the 'alternative' interpretation of a program depends upon its accidental features (T3.34). This develops a peculiar tension in program design that is hard to keep stable, particularly with respect to the informal, and often undefined, mechanism which links the program names with the user's domain. As noted above, the 'objects' that are usually chosen to be referenced in the informal analysis of the problem domain do not normally have all the features required of Wittgenstein's objects.

The Paradigm Leap

The *Tractatus* seems to be an effective description of how programming languages should be linked to a computer in a way that makes the programs have 'sense' (as with meaning) through assignment. There is no problem with the engineering necessity of this approach to sense and meaning. Further, it leaves the human out of the equation. On a broader scale it also sidesteps many of the paradoxes of the linguistic philosophy of the day. However, it has *one fatal flaw* when applied to the human use of language and its author eventually exposed this flaw. He noted that *it is not possible to unambiguously describe everything within the propositional paradigm*. He found that the normal use of language is riddled with exemplary concepts that cannot be bounded by logical statements that depend upon a pure notion of referential objects. One of his illustrations is an attempt to define a game (P69 – P71). Such a definition cannot be achieved that will either exclude all examples that are not games or include all examples that are. It is through such considerations that Wittgenstein proposed a new linguistic philosophy

that was based upon what we (the authors) are calling inferential semantics[105].

It is because of this observation by Wittgenstein that we make the distinction between rational and irrational sets. *An **irrational set**[106] is where no finite set of rules can be constructed that can include unambiguously any member of that set and, at the same time, unambiguously exclude any non-member of that set.*

In contrast to this it is a necessary requirement that formal systems such as the Tractatus are to be based upon a 'rational' set. *A **'rational' set** is a set where there is a finite set of rules that can include unambiguously any member of that set and unambiguously excludes any non-member of that set.* It is a necessary requirement that membership of a set is given as a set of propositions (T3.142). A rational set can then have assigned the value True unambiguously to all its members (T4.3, T4.4, T4.41, T4.43) and False to all its non-members. All the sets referenced by the Tractatus must be rational otherwise such an assignment would be ambiguous and deduction would not be possible (T5.11-T5.123).

By way of illustration of an irrational set consider the set of chairs and a possible specification (figure 3.1). It is always possible to find some exception to a finite set of rules that attempts to identify a member of the set 'chair'. Even if every exception were added to a membership list this would break down by simply discovering a context in which at least one member would cease to be identified as a member through the use of the rules. The more extreme-case additions made to the set, the more opportunities there will be for finding situations that exclude accepted

105 David Gooding (University of Bath, private communication 2004) notes that "The view epitomised by Wittgenstein's Philosophical Investigations is that meaning, grammar and even syntactic rules emerge from the collective practices (the situated, changing, meaningful use of language) of communities of users."

106 The idea of rational and irrational sets was proposed by Jan Townsend Addis (private communication February 2004) who related the irrational sets to Cantor's (1845-1918) irrational numbers. In the case of rational numbers the rule was a member number could be expressed as a ratio of integers. Examples of irrational numbers are $\sqrt{2}$ and π. There are infinitely more irrational numbers than rational numbers. However, as for irrational numbers an irrational set can always be approximately represented by a rational set.

members of the set. We are in a position where most things are not potentially unambiguously describable.

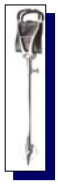
Attempts at providing a rational description of irrational sets has stimulated extensions to the ‘crisp’ set by assigning a ‘value’ to a membership. Examples are fuzzy and probabilistic membership assignments. However, fuzzy sets are rational in that members are assigned a membership number that is explicit and essentially ordinal (T4.464, T5.15-T5.156). Such assignments can be expressed by a finite set of rules. Similarly, a probabilistic assignment of a member is also rational where a rule is in the form of a ratio of integers that specifies its membership (T5.15). None of these mechanisms can be used for irrational sets because the determination of membership is not only dependent upon an individual’s development in the way the world is seen it also varies between individuals because of their different experiences and purposes. Purpose is important because the way the world is viewed can depend upon what is required to be ‘seen’ by the subject. A current example arises as to the definition of what is a ‘planet’.[107] In the past a planet was any object in the sky that moved with respect to the backdrop of stars. In this early case the sun and the moon were considered planets and as such they still play an important role in Astrology. Astronomy required that planets should be bodies that circle any star. The sun was reclassified as a star but planets were to exclude asteroids in the asteroid belt and comets that went outside the solar system. Now some astronomers have excluded Pluto, which is smaller than our moon. There will, of course, be no real termination of this definition while theories of our Universe are still developing.

107 The Committee on Small Body Nomenclature of the International Astronomical Union is due to report on this late in 2006. See the IAU website at http://www.iau.org/HOME.2.0.html

Chair Specification 1:
Designed specifically to be sat upon
Stands on its own
Has four legs
Has a back
Sitter's feet touch floor

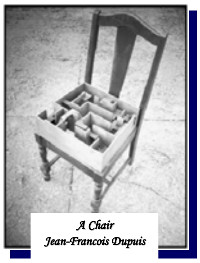
Chair Specification 2:
Designed specifically to be sat upon
Stands on its own
Has four legs
Has a back

Chair Specification 3:
Designed specifically to be sat upon
Stands on its own



Chair Specification 4:
Designed specifically to be sat upon

Chair Specification 5:
Designed to be sat upon



Chair Specification 6:

Figure 3 An attempt at identifying a chair

A more dramatic and poignant example of how a difference in a definition can result in a divergence of action can be seen from the report on the execution of Paul Hill at Starke, Florida, CBSNEWS.com, September 4th 2003. This illustrates how the diverse definitions of 'human' life relating to a foetus can result in dramatically different assessments of what is happening. The report states severally:

> "The execution of Paul Hill for the murder of a doctor who performed abortions and his bodyguard left U.S. abortion providers anxious – and wary that the former minister may become a martyr to the anti-abortion cause and spur others to act violently."

> "Paul Hill's final statement ….. If you believe abortion is a lethal force, you should oppose the force and do what you can to stop it"

> “Paul Hill should be honoured today, the abortionists should be executed. said Drew Holman”

> “We think that unborn children should be protected and it should be law. Said Sheila Hopkins, a spokeswoman for the Florida Catholic Conference. We definitely reject his statement that it was justifiable homicide.”

The abortionist perceived he was doing a public service by removing what he sees as a growth, such as a cancer, that will ruin a woman’s life. Paul Hill (and others) saw the foetus, no matter how young, as potential human life that has been terminated. There is no logical argument that can be applied in this case. The only way minds can be changed is through a slow shifting of views. One mechanism to shift a view is to use tropes that will chart a route from the current view to a new view (e.g. foetus is a cancer or it is a future human life).

Logic is not abandoned by accepting the concept of an irrational set/concept. Even though there are irrational sets we still have rational sets and so denotation remains one mechanism for relating meaning to a name. For irrational sets there is an additional and more important mechanism for meaning assignment based upon human usage and context. It is this latter mechanism that provides the link between the program and the world it is designed to represent and is the other half of the dual semantics.

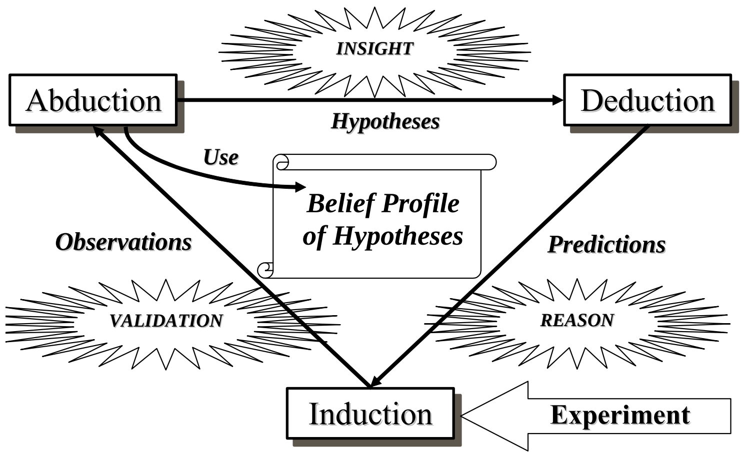


Figure 4. The Peircian Abductive Loop

We can interpret this extended definition of meaning to imply a *process* of inference involving abduction (of hypotheses or new beliefs), deductive testing of hypotheses, and inductive comparisons of predictions with the world (see Figure 4). This process is implemented in Gooding and Addis [Addis & Gooding 1999, 2004] within the framework of scientific practice. During conversation a process is going on where a model of the meaning of words is being constructed mutually, through inference. These models have been labelled construals [Gooding 1990, Arrighi and Ferrario 2004] and boundary objects [Gorman 2004, Bowker and Star, 1999]. This is a *group* activity that constructs something common in the way language and the world may be perceived so that communication can occur. However, these models of word-meanings are only understood through their ability to make predictions and their coherence within a group-dynamic situation. Meanings cannot be observed directly since they only exist within an individual. The hidden dimensions of the model that express concepts are likely to be different for different people, so we have the effect of distinctions having variable boundaries that are defined by examples, not by logical or semantic rules.

The irrational set (or concept) identifies that people work within a dynamic ontology where objects may be subject to changing set boundaries, where an entity can exist across boundaries, and has to so

exist for the dynamism to function. The human mind has no difficulty grasping the concept that all definitions are temporary and may change with the conversational context. Most important of all, new concepts may be created on demand, something no machine could contain within a fixed frame.

We identify the problem of the machine's limitations to achieve human intelligence or interact smoothly with people as less that the machine is limited in its logical processing capability, always assumed to be the root problem in the use of computers in human affairs, than that the human mind is incapable of operating exclusively within the rational set domain. This is because the human mind cannot possess, even potentially, infinite logical power, nor can it contain the whole world of facts (T1.1). It is because of this that people will adopt a temporary rational view to make deductions; a view that approximates their current knowledge of the world.

Predictions

So we have computer programs with a semantics based upon computer bits but we create programs that cannot rationally be assigned meaning to the very problem domain for which they have been written. Programs must remain in the domain of rational sets if they are to be implemented on a machine. However, we do have the freedom to use the program's accidental properties without affecting the program's formal meaning with respect to the computer. We can chose the names we use and select the computer organisation from the possibilities bounded by the essential program.

A proposition, and hence a program, can adopt many equivalent forms. It is the job of a compiler to make a transformation of a program in order that it is acceptable for a particular computer to run it. For any particular computer there are an infinite but bounded number of possible structural forms for a given program. The possibilities are bounded by the limitations of the compiler and the intended final form of the program (the essential program). Apart from these limitations the choice of form is in the hands of the programmer. This means that:

- automatic reverse engineering is impossible unless domain information is used under the guidance of a person. If guidance is used then it can only be semi-automatic. This is because the only human independent meaning of a program is derived from its mapping to computer bit states. Structure and names remain arbitrary and tautological. The only clue as to the meaning of these features is found in the mind of the designer and only another human can, at the moment, make this inferential step.
- design methods will generally only limit what is possible to implement unless they are 'complete'. A 'complete' method is one that constrains the possible designs to that of the limits of the machine being programmed. This is because most methods try to provide a constrained route to a solution but such constraints are derived from limited sets of domains such as business databases.
- machine mismatches can be detected through tautology. This is simply the process of having check bits associated with a set of bits. Such a check is derived from a uniform tautological proposition applied independently of the program. However, this view does open up the possibility of having many different tautological propositions that are orientated towards different machine contexts.
- a general purpose programming language will always be difficult to produce simply because the semantics of machines are independent of the problem domain. Thus any language that is produced will need to have a mapping that makes sense for all possible computers. Further, the problem domain, if it is to do with the human view of the world, is indefinable.
- programs on 'quantum' computers are bounded by operations that do not depend upon knowing an interpretation. Such machines might be ideally suited for irrational computation.
- formal 'objects' (e.g. Windows in OOP) will be stable but informal 'objects' (e.g. persons, life, chairs or games) will never be fully captured or be stable because they are irrational sets.
- it will not be possible to completely represent certain human functionality such as natural language understanding on any machine that is not adaptable.
- increasing a training set for machine-learning algorithms will eventually cause a degradation in its recognition performance if the

set includes irrational concepts (sets). This is because the greater the training set the greater the possibility of eventually having a contradiction.

Some solutions to these issues are that if the context changes then the human must feed back to the machine the fact of contextual change. The effect upon the machine should be that it accepts the human requirement and moves its frame of reference, much as a film director uses a director's viewfinder to frame the scene to be shot. The machine should select out just the requirement of the moment [Visscher 2005]. Some initial attempts to create a program along these lines have been attempted [Stevens et al 2006].

Further, the machine needs to contain the ability to shift its definitions without departing from its inbuilt logic. It might do this by reacting to human language clues as to the current requirement. The machine does not need intelligence as such. It needs rapid adjustability of its whole definition framework. The logical machine could emulate the 'irrational' human by building a patchwork of 'rational' frames only one of which is ever the current frame.

Inferential Semantics

From an engineering point of view the only information that can be experienced by an individual is the result of the interaction of the individual's sense organs with the world. This is not a passive view since these organs are also controlled by an inference engine; namely the human mind. It is only through inference and the senses that we experience the world and relate to other people. So, like the computer, we might be able to trace the sense of our understanding of the world through the tracing of internal constructs to our senses. However, this would not be of any great help to other people since it is unlikely that we are identical in the same way as two computers, constructed according to a defined engineering diagram, are identical. If we were to be different by as little as one bit we could not ever be sure that a 'program' would mean the same if 'run' in different heads or that it would even 'run' at all. So

tracing and knowing the ‘program’ (or our internal constructions) is not very useful.

What could work, from a purely pragmatic point of view, is if individuals could construct models of the world, and other people, that were sufficient to meet the needs of surviving in the world and with others. This model does not have to be exact, just sufficient. However, to do this we have to extend our semantic model to have another definition of meaning; a definition that does not depend upon the direct referencing of objects. For Wittgenstein, the *meaning* of a word was also defined as its *use in language* (P43). As he says:

> “For a *large* class of cases – though not all – in which we employ the word ‘meaning; it can be defined thus: the meaning of a word is its use in the language. And the *meaning* of a name is sometimes explained by pointing to its *bearer*.”

We can interpret this extended definition of meaning to imply a *process* of inference. During conversation, both observed and participating actively, a process is going on where a model of the meaning of words is being constructed through inference. This is a *group* activity and one designed to construct something common in the way language and the world may be perceived; a way that allows communication to occur. However, these models are only understood by their effectiveness, their ability to make predictions and their coherence within a group-dynamic situation. They can never have been ‘seen’ directly since they only exist within an individual. It is the hidden dimensions of the model that express concepts and since these dimensions are likely to be different for different people we have the effect of distinctions having no proper boundaries that can be logically defined.

This lack of boundaries for concepts is the *family resemblance* effect detected by Wittgenstein and illustrated by his example (P67). It is an effect that fuzzy sets, in some cases probability and belief networks, were intended to overcome (see also P71) without losing the power of referential assignment and the power of deductive inference. Very recently a research team in Mexico, in conjunction with Salford University, have started to explore the use of family resemblance with a

learning system in order to approach human performance in categorization [Vadera, Rodriquez & Succar 2003].

The tension caused by the dual semantics that pivots on the essential and accidental meaning of the signs used in programs has been recognised as can be seen by the continued search for new languages, program structuring and systems design methods (e.g. Java, conceptual modelling and object orientation). The central problem of the human context has also been addressed through the pursuit of natural language understanding, naïve physics, case-based reasoning and adaptive interfaces. There is a belief that given sufficient power or moving beyond the Turing machine would somehow solve the problem. This has not been demonstrated by the many-fold increases in computer power or parallel mechanisms such as neural nets. Perhaps those new aims discussed elsewhere [Stepney et al 2005, 2006] may prove successful. However, none of the approaches tried so far have really succeeded. Many of the pursuits have been constrained by the formal bounds represented by the Tractatus, and for those approaches that have broken away, they have not bridged the gap identified here.

The Ethics of a Rational View

There are also social consequences of the view adopted by the Tractatus in that it is assumed that rules can be created for all situations and as such these rules can bypass human judgement. It also assumes that there is only one correct way of seeing the world and so human existence can be governed by some finite set of laws.

There is this important dimension to technology that involves issues beyond the technical. If the 'rational' approach, which dictates that there is only one truth and that as such we can all be bounded by a single 'true' view of the world, is really correct then the need for such things as juries or judgment can eventually be dispensed with and all our behaviour can be assessed through some set of complex rules or laws.

We are not saying that we can impose our will on the way the world 'is'. We do suggest that we have a choice on how we choose to describe it. It is apparent that neither language, nor mental capacity, nor senses, allow us to ever have access to the 'truth' as imagined by the ‘rational’ approach. This is because we are limited and we have to accept this limitation and then govern our actions with that knowledge; we just have to do the best we can.

It is because there is a tendency to support a ‘rational’ view that we now have all the measures of performance and rules of assessment in the modern work environment[108]. If we restrict our ability to use our judgement and limit our responsibility for our own actions through rules as implied by the rational view then it makes our world over constrained. The UK legal system and the way we are assessed as people at work seem to more and more depend upon this view. This is a tragedy since it is attempting to bind us to an unbending framework of a reality which cannot be, in principle, changed.

Currently, such a 'rational' view is pushing the government in the UK down the road of abandoning juries (to save money and time), dictating the way judges give out punishments for crimes (to be fair across the country), arming the police (not a comfortable prospect in the UK), placing draconian rules for motoring and other activities (health and safety) and laying down suffocating measures of assessment on our research and teaching (to ensure standards). The Universities are no longer a haven for exploring new ideas since we are all being pushed by the government to respond to the rational and economic imperatives.

What we suggest is that the 'irrational’ concept, or set, better characterises the world we live in as human beings. This does not mean that we should abandon scientific method or theories but that the theories and methods may have to be imbedded in a dynamic reasoning system as suggested by ‘irrational’ sets. A characteristic of an ‘irrational’ concept (set) is that it can either have a simultaneous disagreement of its bounds amongst a group of people - it is a source of argument - and/or its bounds

---

108 It was this rational view that was the driving force behind Artificial Intelligence during the 1960’s and was the major reason for the demise of Cybernetics as a serious science.

will change over time. This explains why ethical issues will also be open to disagreement and change in the same way.

New Computations

The problem with computers is that you cannot argue with them. Although expert systems will give a line of reasoning as to why it comes to a certain conclusion the line of reasoning cannot be changed by simple discussion, however, 'irrational' sets do allow such argument and a modification of a conclusion. This is also the case with possible alternatives to Wittgenstein's family resemblance. One such alternative is Lakoff's [Lakoff 1986, Lakoff & Johnson 1980] use of prototypes (paradigms) and metaphor instead of reference. We have already concluded that metaphor is a central tool for exacting such change. With either route we have a more acceptable approach to human relationships in that there will always be a need for human judgement because what is acceptable behaviour or performance is a time sensitive and socially dependent notion.

The requirement to encapsulate a wide range and ever changing perceptions of a problem domain within a computer program will be the need for a continuous link with human activity. Such perceptions cannot be predicted and hence planned for in advance. So many of the current principles of design will have to be shelved and two distinct design paths will need to be forged that involve the two independent elements of a program; the formal rational and the informal irrational (figure 5).

The challenge we face is, can we reconstruct computing based upon family resemblance rather than sets, paradigms rather than concepts, and metaphor rather than deduction? Can we devise systems that have judgement rather than decisions? One possibility is that we might be able to write dynamic, socially sensitive interfacing-compilers that can match any program to any user (see figure 5).

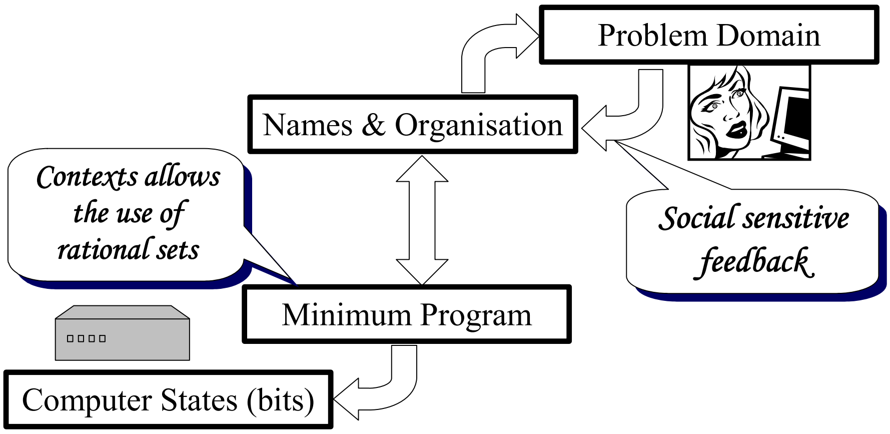


Figure 5. Showing where change can occur to solve the dual semantic problem

Such a compiler would be in 'conversation' with its user, other users and machines via (say) the Internet absorbing the human cultures and language so that its generated semantic and semiotic mappings make a program usable by a person. This will provide a more natural communication between people and machines; it may identify what is really meant by common sense.

Conclusions

Irrational sets or concepts are not technical solutions, as might be Fuzzy Sets, Probabilistic Reasoning, Genetic Algorithms, Neural Nets, and so on. But it *is* a way of looking at the world when considering the *use* of these kinds of solutions. It forces a link between different informatics issues such as HCI, software and hardware architectures and domain analysis. It exposes the limitations of certain research paths that have been taken to achieve human emulation. In particular it implies that formal approaches such as theorem proving, expert systems and natural-language understanding, when based upon a fixed semantics, are doomed to failure. What we found surprising was that it also has a strong ethical aspect that goes beyond technology and impacts upon meta-ethical reasoning.

References

Addis T. R. & Gooding, D. C. 1999 Learning as collective belief-revision: simulating reasoning about disparate phenomena, Proceedings of the AISB'99 Symposium on Scientific Creativity, ISBN 1 902 956044.

Addis T. R. & Gooding, D. C., 2004, *Methods for an Abductive System in Science*, Part of the Abduction and Creative Inferences in Science Symposium, led by: Lorenzo Magnani, CogSci 2004, 26th Annual Meeting of the Cognitive Society, August 5-7.- **Publication Pending**

Addis T. R., Townsend-Addis J. J., Billinge D., Gooding, D. C. & Visscher, B., 2004, *Tracking Irrational Sets*, Model-Based Reasoning in Science and Engineering, Abduction, Visualization and Simulation, MBR'04, Pavia, Italy, December 16-18. - **Publication Pending**

Addis T. R., Visscher B., Billinge D., Gooding, D. C. 2005 Socially Sensitive Computing Grand Challenge in Non-Classical Computation - International Workshop, University of York 18-19 April, http://www.tech.port.ac.uk/staffweb/addist/MS_GC7_05_SSC.pdf

Arrighi, C. and Ferrario, R., 2004 , Abductive reasoning, interpretation and collaborative processes, in *MBR04: Model-Based Reasoning In Science And Engineering, Abduction, Visualization, and Simulation*, University of Pavia, December 2004.

Billinge, D. & Addis, T. R. 2004 *Music to our ears: a required paradigm shift in Computer Science,* European Conference on Computing and Philosophy, University of Pavia, Italy

Bowker, G. C and Star, S. L. , 1999, *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press

Dreyfus, H.L. 1992 *What computers* still *can't do*. MIT Press, Cambridge Mass ISBN0262041340

Gooding D. C. 1990, *Experiment and the Making of Meaning*. Dordrecht and Boston: Kluwer Academic.

Gooding, D. C. & Addis T. R., 2004, *Modelling Scientific Experiments as Mediating Models*, Model-Based Reasoning in Science and Engineering, Abduction, Visualization and Simulation, MBR'04, Pavia, Italy, December 16-18. - **Publication Pending**

Gorman, M. E., 2004, *Trading zones, moral imagination and socially sensitive computing* – Model-Based Reasoning in Science and Engineering, Abduction, Visualization and Simulation, MBR'04, Pavia, Italy, December 16-18. - **Publication Pending**

Kleene, S.C. 1967. *Journal of Mathematical Logic*, New York: Wily.

Lakoff G. 1986. *Women, Fire, and Dangerous Things*. University of Chicago Press.

Lakoff, G & Johnson, M. 1980 *Metaphors We Live By*, Chicago University Press ISBN 0-226-46801-1

Luce D. & Raiffa H, 1967 *Games and Decisions: introduction and critical survey* 7th ed. J. Wiley & Sons, NY

Smith B. C., 1982 *'Linguistic and Computational Semantics'*, 20th Annual Meeting of the Association for Computational Linguistics, 15-18 June, University of Toronto, pp9-15.

Stepney S., Braunstein S. L., Clark J, A., Tyrrell A., Adamatzky A., Smith R. E., Addis T. R**.**, Johnson C., Timmis J., Welcj P., Milner R. and Partridge D, 2006, *Journeys in Non-Classical Computing II: A Grand Challenge for Computing Research* International Journal of Parallel, Emergent and Distributed Systems, Vol 21, No 2, April ISSN 1744-5760, Taylor & Francis pp97-125.

Stepney S., Braunstein S. L., Clark J, A., Tyrrell A., Adamatzky A., Smith R. E., Addis T. R**.**, Johnson C., Timmis J., Welcj P., Milner R. and Partridge D, 2005, *Journeys in non-classical computation I: A grand challenge for computing research*. International Journal of Parallel, Emergent and Distributed Systems, Vol 20, No 1, April ISSN 1744-5760, Taylor & Francis pp5-19, March

Stevens C. J. C, Visscher B-F, & Addis, T. R., 2006, '*Automating the Abductive Inference Loop*' **Accepted** Paper ID: 01525, 2006 IEEE International Conference on Systems, Man, and Cybernetics.

Vadera S, Rodriquez A, Succar E 2003 *Family resemblance, Bayesian networks and exemplars*. AISB Quarterly No.114 p.1 & 11

Visscher B. 2005 Exploring Complexity in Software Systems, PhD Thesis, University of Portsmouth http://www.tech.port.ac.uk/staffweb/addist/amuse.html

Wegner P. & Eberbach E. 2004 *New Models of Computation*, The Computer Journal, Vol 47, No1, pp. 4-9.

Wittgenstein L 1921 *Tractatus Logico-Philosophicus* edition 1961Routledge and Kegan Paul London

Wittgenstein L 1953 *Philosophical Investigations,* Blackwells Oxford

COMPUTATIONALISM AND THE LOCALITY PRINCIPLE

David Longinotti
Department of Philosophy
University of Maryland
(301) 483 4766
longinotti@hotmail.com

Computationalism, a specie of functionalism, posits that a mental state like pain is realized by a 'core' computational state within a particular causal network of such states. This entails that what is realized by the core state is contingent on events remote in space and time, which puts computationalism at odds with the locality principle of physics. If computationalism is amended to respect locality, then it posits that a type of phenomenal experience is determined by a single type of computational state. But a computational state, considered by itself, is of no determinate type – it has no particular symbolic content, since it could be embedded in any of an infinite number of algorithms. Hence, if locality is respected, then the type of experience that is realized by a computational state, or whether any experience at all is realized, is under-determined by the computational nature of the state. Accordingly, Block's absent and inverted qualia arguments against functionalism find support in the locality principle of physics. If computationalism denies locality to avoid this problem, then it cannot be considered a physicalist theory since it would entail a commitment to phenomena, like teleological causation and action-at-a-distance, that have long been rejected by modern science. The remaining theoretical alternative is to accept the locality principle for macro events and deny that formal, computational operations are sufficient to realize a phenomenal mental state.

Computationalism, a specie of functionalism, is the thesis that a mind is essentially a computer. Proponents of this theory claim that it

offers a physicalist account of mental states that provides for their multiple realizability.[1] However, a number of objections have been raised against functionalist accounts of phenomenal experience. Prominent among them are Block's absent and inverted qualia arguments, which rest on his intuitions.[2] In this paper, I will argue that Block's intuitions have a scientific foundation in the locality principle, a basic tenet of physics that is known to philosophers as 'supervenience'. As currently formulated, computationalism defies locality, and thereby entails a bizarre physics. If it is modified to respect locality, then computationalism under-determines the type of phenomenal experience it claims to define.

By "under-determination" I mean the inability of a theory to account for properties of the phenomenon it is intended to describe. This sort of under-determination is the inverse of Quine's. He posits that, for any set of data, there are countless theories, none scientifically better than the others, that are consistent with that data.[3]

For purposes of this paper, I will adopt Kim's view that an event is a set of properties had by an entity at a time,[4] and will use the terms "event" and "state" interchangeably. I employ "qualia" to refer to the sort of properties that distinguish one type of phenomenal experience from another.

Functionalism and Computationalism

Functionalism is the view that a mental state is defined by its causal relations with sensory inputs, behavioral outputs and other mental states. Functionalist theories come in many varieties, but can be classified into two main types. These are generally labeled, after Block, as Functional State Identity Theories (FSIT) and Functional Specification Theories (FSP).[5]

Although FSIT and FSP both rely on Ramsification, the result of that process is viewed differently by the two theories. Ramsification produces a definition of a mental state-type, M_J, like the following:[6]

$$M_J x \leftrightarrow T(P_1 x \;\&\; P_2 x \;\&\; \ldots \;\&\; P_N x) \;\&\; P_J x \qquad (1)$$

where T is a set of predicates describing causal interactions among the theoretical and observational entities represented by the P's. So x has mental state M_J if and only if x has a dispositional character that can be mapped to theory T, and x is now in the physical state corresponding to the predicate P_J within T. The state represented by P_J is termed by Shoemaker the 'core realizer' of M_J. For example, if M_J is pain, it might be core-realized by c-fibers firing.

FSIT theorists take (1) to be a *constitutive* analysis of M_J.[7] The causal interactions represented by T are seen as ontologically essential for M_J, so M_J is not realized unless P_J occurs within a causal network that satisfies T. For FSIT, the realization of a quale by the core event P_J is a *contingent* fact; P_J and M_J are *relational* properties.

For FSP theorists, T provides a kind of *conceptual* analysis of M_J; it is the means by which we name (i.e., 'pick out') a mental state-type. FSP maintains that the relationship of M_J to T is analytical, and that we empirically discover the specific physical event, P_J, that fills the causal role of M_J in T. If we learn that it is 'c-fibers firing' that occupies that role then, by the transitivity of identity, it is logically *necessary* that P_J is identical with M_J.[8] In FSP, M_J and P_J are *intrinsic* properties.

Computationalism is a type of FSIT functionalism; it views mental

---

states as relational in nature and asserts that the defining causal relations in (1) are correctly characterized as formal operations on symbols. A specific mental state is said to be realized by a particular computational state within a set of computational operations, that is, within a specific algorithm. Computationalists differ in their views of the type of theory represented by T in (1), but there is wide agreement among them that a Turing machine is capable of executing any algorithm described by T.

A 'Turing machine' is a hypothetical computational mechanism envisioned by Alan Turing that operates, in part, by detecting the presence or absence of a mark on a tape. Based on the outcome of that detection, the current state of the machine and the rule of operation accompanying that state, the mechanism produces (or erases) a mark on the tape and moves the 'read/write' component to a new tape position, one unit to the left or right.[9] Church and Turing posited that any intuitively computable function can be computed on such a machine.[10] Computationalism entails that, given the proper inputs from sensors and the right sequence of computations, a Turing machine would be able to experience any type of quale.

Multiple realizability is a straightforward corollary of computationalism. In principle, a computer with the capabilities of Turing's machine can be mechanized with any material having a physical property that can persist and be changed, such that it can be used as a symbol. Computationalists see this as a key virtue of their theory because the reliance on such 'blind' phenomena obviates any need to postulate a homunculus, some 'little person' in the mind who receives the processed input from the senses as though he were watching an inner television. Instead, the process for realizing a mental state can be broken down into simpler and simpler steps until each step can be implemented with a very simple mechanism. Rey says that "what does the work of the homunculus is simply brute physical causation."[11]

The Metaphysics of Qualia

Qualia realism is the view that phenomenal experiences are metaphysically distinct entities, and that qualia are their properties. Shoemaker is an FSIT functionalist who is also a qualia realist.[12] In contrast, Rey is a computationalist who is an eliminativist concerning qualia.[13] Like Dennett, he maintains that nothing in the world answers to the properties that qualia realists ascribe to phenomenal experiences.[14]

The locality argument against computationalism does not depend on qualia realism. What it necessary for the argument is that there at least *seem* to be properties, real or illusory qualia, that have the

determinate nature ascribed to particular experiences. To my knowledge, there is no philosopher who denies this.

Locality and Supervenience

The locality principle of physics is the denial of 'action-at-a-distance'. Healey calls it 'spatiotemporal separability', and characterizes it as the view that "any physical process occupying spacetime region *R* supervenes upon an assignment of qualitative intrinsic physical properties at spacetime points in *R*."[15] In other words, an event is determined by properties that are spatially and temporally collocated with that event. The locality principle can be seen as a partial expression of Einstein's special theory of relativity, which posits that nothing remote from spacetime location (z, t) can have an influence at (z, t) unless it transmits something (at no more than light speed) that is present at z at time t. Einstein's theory supplanted Newton's hypothesis that the gravitational effect from a body like the sun reaches instantaneously across space to exert a force on other bodies.[16]

In the quote above from Healey, he uses the word "supervene" as part of his statement of the locality principle. Locality is implicit in the philosophical concept of supervenience, as in this formulation from Kim:

> Mental properties *supervene* on physical properties, in that necessarily, for any mental property M, if anything has M at time t, there exists a physical base (or subvenient) property P such that it has P at t, and anything that has P at a time has M at that time.[17]

Supervenience is a corollary of locality. If what is realized at (z, t) is determined by properties at (z, t), as locality states, then things with qualitatively identical, physical base properties, P, must realize identical mental event-types, M, as supervenience holds. The supervenience of the mental on the physical is considered by many to be the minimal commitment to physicalism.[18] In what follows, I will use 'supervenience' and 'locality' interchangeably, to mean that the occurrence of an event at (z, t) depends only on properties at (z, t).

There appear to be some events at the quantum level that exhibit non-locality.[19] However, any such events are insignificant for the sort of large-scale 'brute' phenomena that, according to computationalists, are sufficient for realizing a mental state. Accordingly, the possible existence of some non-local phenomena is not relevant to computationalism's claims.

Turing respects the locality principle in that the next state of his hypothetical computer is determined only by the current state, and by the set of rules associated with that state. These rules are embodied in the mechanism that changes the computer from one state to the next, such that only a local, causal influence is producing an effect at any time.

Block's Arguments Against Functionalism

Block argues that functionalism suffers from problems concerning 'absent qualia' and 'inverted qualia', among other difficulties. With regard to absent qualia, Block asks us to imagine that the citizens of China have each been provided with a two-way radio, and that they operate the radios in a manner that mimics the neural interconnections of a conscious human brain. That is, the causal interactions realized by the operation of the radios satisfy the functional algorithms that are said to be sufficient for a series of mental states. But Block is doubtful that, as a result of the radio communications, the nation of China would comprise a conscious entity.

The inverted qualia argument questions whether the specific character of a qualitative state is functionally definable. Per Block, "it seems that we could be functionally equivalent even though the sensation that fire hydrants evoke in you is qualitatively the same as the sensation grass evokes in me."[20] If Block is correct, then computationalism is a theory that under-determines the qualitative nature of a particular phenomenal experience.

The Locality Problem for Computationalism

Imagine that a Turing machine is in the 'core' state minus one (P_{J-1}), of the presumably long and complex algorithm for realizing the taste of a vintage bordeaux. Assume, further, that the next operation in the algorithm is to mark a logical "1", move the tape one unit to the right, and put the machine into state P_J. When this step is completed, the machine is said to taste bordeaux. But, by the locality principle, the mechanism that is the Turing machine is influenced only by local properties. Metaphorically speaking, it does not 'know' how it got to P_{J-1}. Therefore, simply putting it in state P_{J-1} and having it execute the next step, to P_J, should suffice for the machine to savor the taste of a fine wine. But computationalism denies this, and so it violates the locality principle.

Amending computationalism to comply with locality has the consequence that computationalism under-determines the type of quale

that is realized. To see this, it is necessary to bring computationalism into conformance with supervenience. Per supervenience, realizing M_J at t depends only on realizing some P_J at t, but computationalism claims that to realize M_J at t, P_J must have the right causal antecedents and consequences. The most direct way to maintain the spirit of computationalism, while respecting locality, is to remove the network of causal interactions represented by T in (1). The resulting functional definition of M_J is

$$M_Jx = P_Jx \qquad (2)$$

where P_J describes a physical event having a certain formal, symbolic meaning. T now has the same function that it does in FSP, namely, to provide the way of referring to the mental state that is being defined. The difference with FSP is that, for computationalism, P_J would not be an intrinsic property.

But (2) is not sufficient to specify a particular quale. Marking a logical "1", for example, could be the core realization event for different qualia having different algorithms as their reference. Being composed of 'brute' processes, the computer does not 'know' (i.e., is not causally influenced by) the algorithm within which it came to mark this symbol. Even more problematic for computationalism is that the symbolic meaning of the physical mark is a relational property; it depends on events that do not exist at the time and place of the event labeled P_J. So, by locality, the symbolic meaning can't influence what the mark realizes. The argument can be summarized as follows:

1. If computationalism is modified to respect locality, then a quale is realized by a single, token 'core' event having a particular symbolic description.

2. Qualitatively identical token events could have different symbolic descriptions, such that they realize different qualia.

C. If computationalism respects locality, then it under-determines the type of quale that is realized by the tokening of a particular event.

That is, if the computer is realizing *any* type of quale due to the token event, it must be realizing *all* types of qualia. And since virtually any event can be given any symbolic description when considered in isolation, virtually any event must be realizing all qualia. So applying

locality to computationalism entails the under-determination of a quale and, with it, a very radical panpsychism.

The eliminativist concerning qualia may see this argument as supporting his position, but that is not the case. The eliminativist acknowledges that there seems to be something with the determinate properties of a quale that is realized by the core state P_J, but maintains that this determinate thing is an illusion. So the illusion has determinate properties, in which case the same locality considerations apply to the illusion as to a quale – a single event having a symbolic description is insufficient, in virtue of that description, to specify the properties of an illusion. This problem could be avoided if nothing ever happened to us that resembled a determinate phenomenal experience but, surely, this is not the case. Accordingly, Block's intuition concerning inverted qualia finds deductive confirmation in the scientific principle of locality. Given locality, functionalism can't account for the determinate nature of a quale, or of the illusion of a quale.

Similar considerations apply to Block's 'absent qualia' argument. Whether or not a quale is realized by an event depends, for a 'localized' computationalism, on the symbolic meaning of that event considered by itself. But the same type of event could have a symbolic meaning that is, or is not, associated with the realization of a quale – or it could have no symbolic meaning at all. So computationalism, if it is made consistent with locality, is inadequate to specify whether or not any quale is realized by an event-type. Accordingly, Block's absent qualia intuition also receives support from the locality principle.

Kripke's argument against functionalism can be seen to rest implicitly on the locality principle:

> [in functionalism] the causal role of the physical state is regarded by the theorists in question as a contingent property of the state, and thus it is supposed to be a contingent property of the state that it is a mental state at all, let alone that it is something as specific as a pain . . . this notion seems to me self-evidently absurd. It amounts to the view that the very pain I now have could have existed without being a pain at all.[21]

His characterization of functionalism as relying on contingent causal roles indicates that Kripke's target is FSIT functionalism, which includes computationalism. Implicit in his view is that pain could exist despite different causal antecedents and consequences, that is, he sees pain as an intrinsic property, one that satisfies supervenience. By itself, Kripke's brief argument begs the question against the functionalist, who would challenge the assumption that the pain could exist without the rest of the causal relations. The addition of the locality principle provides Kripke

with the premise needed to support his claim that pain could exist without the surrounding causal interactions, and thereby makes his argument against functionalism sound.

Given locality, the only apparent theoretical recourse is to hold that the existence and specific nature of a particular quale (or the illusion thereof) depend on the type of material on which the realization event supervenes - in which case multiple realizability is false for phenomenal mental states.

A number of objections might be raised against the locality argument. In what follows, I reply to some actual and anticipated criticisms.

Objection: Causal Chains and the Transitivity of Causation

One response that has been made to the locality argument is an appeal to the transitivity of causation: the Turing machine would not have reached the core state P_J if it hadn't executed all the prior steps of the computational sequence – in which case the prior events in the causal chain are necessary for P_J. But this objection confuses *tokens* of states with *types* of those states. It's true that a *token* event, like a particular instance of a Turing machine reaching P_J, is the result of a particular sequence of events. But this is not true of an event considered as a *type*, and what computationalism is intended to define are mental state-types.

By way of analogy, any *token* opening of the door to my home is the result of a particular causal chain that goes back to the origin of the universe. But the *types* of effects that result from a particular opening are not influenced, for example, by the way I got to the door. If I open the door with the same force on two successive evenings, and if the door has not changed in the interim, then the types of effects on the door at the two openings will be exactly the same.

There is a legitimate use of the words "cause" and "because" that conveys a kind of historical meaning, as in 'he opened the door quickly because it was raining.' That is, the rain was one link in a token causal chain that resulted in a particular door opening. But the *types* of effects realized in the door do not depend on the rain; they depend only on the type of force applied to the door.

Similarly, the *types* of effects produced in the core state, P_J, of a Turing machine implementation of T result only from the types of physical properties involved in the transition from P_{J-1} to P_J. They are not influenced by the particular causal chain that resulted in the tokening of state P_{J-1}.

Objection: Parallel Computation

Another reply has been that, since the *sequence* of operations contributes to the locality problem, the problem can be removed by executing the algorithm in a parallel fashion rather than serially, as a single Turing machine would. But a parallel implementation would simply shift the locality problem from the temporal dimension to the spatial.

Functionalism is the view that the constitutive events of a mental state are causally related. Even if there were a large number of machines, each doing just a small segment of the computation, there must be something that causally links the machines. So suppose some simple, causal mechanism is implemented that detects the completion of all the sub-calculations. This could be a circuit that realizes a logical AND function. Each machine sends a logical "1" to the AND gate when it finishes its portion of the computation. When all the "1's" are present at the inputs to the AND circuit, an output "1" is generated and, presumably, the machine experiences the relevant quale.

But this doesn't help against locality. The AND circuit is a brute device; there is no homunculus in the AND circuit that 'sees' the distant events that caused the inputs. Just as no type of change in the Turing machine is influenced by what happened in the past, no event-type in the AND circuit is influenced by what transpired in remote locations. The type of event that is the output of the AND circuit depends only on the local inputs; the same types of inputs could have come from many different kinds of sub-calculations. So, once again, if a quale is realized by the AND circuit, then the type of quale that is realized must be under-determined.

Objection: Add a Memory

An objection might be made that these problems can be overcome by adding a bit more functionality to the Turing Machine. The computer could simply keep a record on the tape of the states that have been executed, and could check this record to ensure that all the steps had been completed before proceeding to the core state P_J. But this will not help either. The record could simply be forged. Since it relies on 'brute' processes, the machine does not 'know' how the 'record' got there. That is, there is no causal mechanism in the machine that makes the next state dependent on the actual process that produced the marks comprising the record.

Objection: Qualia as a Computational *Process*

A defender of computationalism might try to avoid these difficulties by identifying a particular phenomenal experience with a computational algorithm, rather than with a state within the algorithm. This would change the functional definition by eliminating the core state. The resulting definition is:

$$M_J x \leftrightarrow P_1 x \; \& \; P_2 x \; \& \; ... \; \& \; P_N x \tag{3}$$

This approach views a quale as a process rather than as a state, but it does not thereby avoid the under-determination problem. The quale must be realized at some stage of the algorithm – with the first state, the last (i.e., the completion of the process), or some state in between. But the physical event that comprises any state could have any symbolic interpretation, when considered locally. So if a quale is experienced at a particular state, then *all* qualia must be experienced at that state – and the under-determination persists.

At this point, a computationalist might appeal to dispositions – in particular, to the disposition of the computer to execute the entire algorithm. She might argue that only if the computer has the proper disposition does it realize the specific quale at some machine state. But this, too, is a denial of locality. It amounts to the claim that what happens at (z, t) is influenced by things that happen (or don't happen) at places and times remote from (z, t) – namely, the manifestation of the disposition. There are arguments against functionalism, by Maudlin and Antony, that are based on the intuition that the dispositions of a machine that are not manifest in a particular process can't influence the effects realized by that process.[22] This anti-functionalist intuition also finds scientific support in the locality principle.

Objection: Relational Properties and Causation

The locality principle entails that relational properties have no causal influence. Geach labeled relational characteristics as 'mere Cambridge properties', those without any power to change things in the world.[23] Some, like Francescotti, dispute this. In support of the claim that relational properties can be causal, he gives the following example:

> Suppose a pilot notices a burning barn, a little while later notices Jack, calculates the distance between Jack and the barn, and forms the belief

> that Jack is fifty miles east of a burning barn. In this case, the pilot's belief is caused, at least in part, by Jack's being located fifty miles east of a burning barn.[24]

But this is not correct if it is intended to mean that the relational property as such contributes to the pilot's belief. What actually occurs is that the pilot locally generates and compares information with regard to the locations of Jack and the barn – perhaps by noting the time it takes, at a certain velocity, to fly from one to the other. Producing this information requires separately receiving energy in some form from Jack and the Barn, to establish when the plane is at each position. Each event of this process is locally determined.

What about an historical property, like 'having been painted by Degas'? Doesn't such a relational property have causal power – the power, for example, to induce a prospective buyer to pay more for the painting? The answer is no. If the buyer believes that the work was painted by Degas, even though it was not, then he will still pay the higher price. It is the buyer's belief that determines his behavior, not the actual history of the painting. And the belief supervenes on local properties.

The Locality Principle Revisited

If the computationalist is not to surrender the basic elements of his theory, it seems that his only recourse is to challenge the locality principle. Some philosophers do hold a non-local view of causation. For example, Taylor argues that if every cause were simultaneous with its effect, there could be no such thing as a causal chain. That is, Taylor denies that effects supervene locally in time on their causes:

> If some event A, for example, causes B, which in turn causes C, which in turn causes D, then in case every cause is simultaneous with its effect, it follows that when A occurs, then the others, and indeed every event in the universe that is in any way causally connected with A, must occur at the same time.[25]

But what Taylor sees as a continuous causal chain is viewed by modern physics as a sequence of discrete causal events tied together by the effects of inertia. For example, consider three billiard balls; ball A contacts ball B, causing it to move toward ball C. In the modern view of causation, B is caused to move by A, but that causal event exists only while the two balls are in contact. Then, B moves in the direction of C but, after B's contact with A ends, there is nothing like a force that is causing B to move. Indeed, as Einstein theorized, there is a frame of

reference in which ball B is not moving at all. When B strikes C there is, strictly speaking, a new causal event, not a continuation of the cause that was A's impact with B. As noted previously, the *type* of event that occurs when B contacts C depends only on the local properties at the time of the impact.

Taylor's kinematics are basically the same as those of Aristotle; they both hold (Taylor implicitly) that something cannot move unless there is a cause that is acting on it. But Galileo overturned Aristotelean physics of motion by positing the existence of inertia, which is a *property* of a body (not a cause acting on a body) such that it resists changes to its kinematic state.[26] That is, a body placed in motion will, without any cause, continue in motion. Any change in motion occurs simultaneously with the application of a force to the body.

Each thing has an infinite number of relational properties relative to other things in the universe. If relational properties had causal influence, then the scientific method would be unworkable. As Einstein expressed in a letter to Born, "If this axiom [i.e., locality] were to be completely abolished, the idea of (quasi-) enclosed systems, and thereby the postulation of laws which can be checked empirically in the accepted sense, would become impossible."[27]

Denying locality would render computationalism an anti-physicalist theory, in that it would entail a commitment to types of phenomena that have been rejected by modern science. These include spatial action-at-a-distance, ghostlike influences on the present from events in the past, and Aristotle's teleological mode of causation, wherein events of the future partly determine present events.

The Generalization Objection

A final objection might be that, if the locality argument were sound, then it would not be possible to have a functional definition of anything. But many things, like carburetors and capacitors, have *only* functional definitions. Therefore, the argument is not sound.

But the first premise in this objection is not correct. The locality argument targets FSIT functional definitions, which are relational in nature. It is not applicable to FSP functional definitions, which depend on intrinsic properties. Block puts the difference between FSIT and FSP this way:

> The functional state identity theorist would identify pain with the [relational] property one has when one is in a state that is caused by pin pricks and causes loud noises and also something else that causes

> brow wrinkling. The functional specifier would define pain as the *thing* that is caused by pin pricks and causes . . [etc.] .[28]

Most of our functional definitions are of the FSP variety. We think that a carburetor can still be carbureting if we arrange things properly on the workbench – its activity as a carburetor does not depend on its relationship to an engine. Similarly for an electronic capacitor or transistor. Such things have functional natures that are determined by their *intrinsic* characteristics.

But FSIT holds that something different happens when c-fibers are locally stimulated within a causal network, as opposed to when the same type of stimulation is applied without the additional causal relations. For FSIT, pain is realized in the first case, but not in the second. FSP says that pain is realized in both cases, if it is realized at all. FSP definitions comply with locality and supervenience, while FSIT definitions do not.

Summary

Computationalism posits that a phenomenal experience is realized by a 'core' computational state within a causal network of such states. This implies that what is realized at the core state is influenced by events remote in space and time, which puts computationalism at odds with the locality principle of physics and with supervenience. Computationalism can be modified to respect supervenience and locality by removing the necessity for a causal network. But then the type of quale that is realized at the core state, and whether a quale is realized at all, would be under-determined by the theory. Denial of locality would render computationalism an anti-physicalist theory since it would entail a commitment to phenomena, like teleological causation, that have been rejected by modern science, and would place computationalism in conflict with the special theory of relativity. The remaining alternative is to accept the locality principle for macro events and deny that formal, computational operations are sufficient to realize a phenomenal mental state.

Notes

[1] For a seminal paper on functionalism, see Putnam (1975).

[2] Block (1980a).

[3] Quine (1960).

[4] Kim (1973).
[5] Block (1980b).
[6] Shoemaker (1981).
[7] For example, Rey (1997), p.29.
[8] Lewis (1980a).
[9] See Turing (1936). Rey (1997) provides a concise description of Turing's hypothetical machine.
[10] See, for example, Copeland (2004).
[11] Rey (1997), p. 267.
[12] Shoemaker (1991).
[13] Rey (1997), pp. 305-308.
[14] Dennett (1997).
[15] Healey, R. (1999), p. 7.
[16] Feynman (1963), Vol. 1, Ch. 15.
[17] Kim (1998), p. 9.
[18] For example, Lewis (1999).
[19] Maudlin (1994).
[20] Block (1980a), p. 288.
[21] Kripke (1972), p. 146.
[22] See Maudlin (1989) and Antony (1994).
[23] Geach (1969).
[24] Francescotti (1999), p. 296.
[25] Taylor (1966), p. 38.
[26] An account of Galileo's revision of Aristotle's physics of motion is provided by Dampier (1936).
[27] As quoted in Maudlin (1994), p. 7.
[28] Block (1980b), p. 180.

References

Antony, M. (1994). "Against Functionalist Theories of Consciousness." *Mind and Language* 9, 105-123.

Block, N. (1980a)."Troubles with Functionalism," in N. Block, ed., *Readings in Philosophy of Psychology, Vol. 1.* Cambridge, Mass.: Harvard University Press, 268-306.

________ (1980b). "What is Functionalism," in N. Block, ed., *Readings in Philosophy of Psychology, Vol. 1.* Cambridge, Mass.: Harvard University Press, 171-185.

Copeland, J. (2002). "The Church-Turing Thesis," in E. Zalta, ed., *The Stanford Encyclopedia of Philosophy (Fall 2002 Edition)*, <http://plato.stanford.edu/>.

Dampier, W. (1936). "From Aristotle to Galileo," in J. Needham and W. Pagel, eds., *Background to Modern Science*, New York: Macmillan and Company.

Dennett, D. (1997). "Quining Qualia," in N. Block, O. Flanagan, G. Güzeldere eds., *The Nature of Consciousness*, Cambridge, Mass.: MIT Press, 619-642.

Feynman, R. (1963). *The Feynman Lectures on Physics.* New York: Addison-Wesley.

Francescotti, R. (1999). "Mere Cambridge Properties." *American Philosophical*

Quarterly 36, 295-308.
Geach, P. (1969). *God and the Soul*. London: Routledge & Kegan Paul.
Healey, R. (1999). "Holism and Nonseparability in Physics," in E. Zalta, ed., *The Stanford Encyclopedia of Philosophy (Winter 2004 Edition)*, <http://plato.stanford.edu/> .
Kim, J. (1973). "Causation, Nomic Subsumption and the Concept of Event," *Journal of Philosophy* LXX, 8, 217-235.
Kripke, S. (1972). *Naming and Necessity*. Cambridge, Mass.: Harvard U. Press.
________ (1998). *Mind in a Physical World*. Cambridge, Mass.: MIT Press.
Lewis, D. (1980a). "Psychophysical and Theoretical Identifications," in N. Block, ed., *Readings in Philosophy of Psychology, Vol. 1.* Cambridge, Mass.: Harvard U. Press, 207-215.
________ (1980b). "Mad Pain and Martian Pain," in N. Block, ed., *Readings in Philosophy of Psychology, Vol. 1.* Cambridge, Mass.: Harvard University Press, 216-222.
________ (1999). "New Work for a Theory of Universals," in J. Kim and E. Sosa, eds., *Metaphysics: An Anthology*, Malden, Mass.: Blackwell, 209-222.
Maudlin, T. (1989). "Computation and Consciousness." *The Journal of Philosophy*, 48: 407-32.
________ (1994). *Quantum Non-Locality and Relativity*. Cambridge, Mass.: Blackwell.
Putnam, H. (1975). "The Nature of Mental States," in N. Block, ed., *Readings in Philosophy of Psychology, Vol. 1.* Cambridge, Mass.: Harvard University Press, 223-231.
Quine, W. (1960). *Word and Object*. Cambridge, Mass.: MIT. Press.
Rey, G. (1997), *Contemporary Philosophy of Mind.* Cambridge, Mass.: Blackwell.
Shoemaker, S. (1981). "Some Varieties of Functionalism." *Philosophical Topics*, vol. 12, 93-120.
________ (1991), "Functionalism and Qualia," in D. Rosenthal, ed., *The Nature of Mind*, Oxford: Oxford University Press, 395-407.
Taylor, R. (1966), *Action and Purpose*. Englewood Cliffs, N.J.: Prentice-Hall.
Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, Series 2, 42 (1936-37), 230-65.

CO-EVOLUTION OF HUMAN AND ARTIFICIAL COGNITIVE AGENTS

LAURA PANĂ

University "Politehnica" of Bucharest. lcpan20032000@yahoo.com

Abstract. Cognitive techniques are now multiplied at each level of the intellectual activity, from the informing, to the creative activities. By common activities in the present new technical cognitive environment, human and artificial cognitive agents are gaining some common characteristics. Cognition forms are themselves evolving in all fields of culture and new forms of cognition appear. Some philosophical techniques are applied in cultural fields as aesthetics and ethics, techniques adapted for new expressions of these philosophical domains, as well as Information Aesthetics and Machine Ethics. Prospective techniques can be used in order to anticipate some trends of the information techniques assisted philosophical thinking. Knowledge groups, groupware and knowledge management problems and techniques are studied. Some cognitive techniques can be common to artificial and human cognitive agents. Applied by humans and machines, who can meet in the middle of the road between the natural and the artificial, these cognitive techniques facilitate a common, faster evolution.

Keywords: cognitive techniques, knowledge groups and groupware, artificial cultural evolution and environment

Human and artificial beings are acting and interacting now in an artificial technical environment which is partially a cognitive environment. Machines cumulate multiple and meaningful functions related to man and society. Artificial agents are created not only for assist, but for replacing humans in processes as fabrication, business, services, communication, research, education or entertainment.
This more and more artificial world, generated by man-machine interaction, produces not just complication of the machine, but of the man himself and of its cultural values. Human species evolves in all its dimensions: biotical, psychical, social and cultural: now it evolves towards artificiality.
Society evolves too, mainly under the influence of the information and management technologies advances. We now have an information society and partially even a knowledge society, but the following desirable step, the building of the culture society, will be based, in our opinion, on **a common culture of humans and machines,** starting from the fact that human and artificial agents are now going to explore and to populate a global cultural environment, structured by common cognitive values and techniques.
The cultural dimension of humans is also transformed and a technical man is born. All types of values are renewed, because of the emergence of new human needs, often satisfied by virtual relations and virtual means, in a virtual environment, which is a technical and mainly intellectual artificial environment.

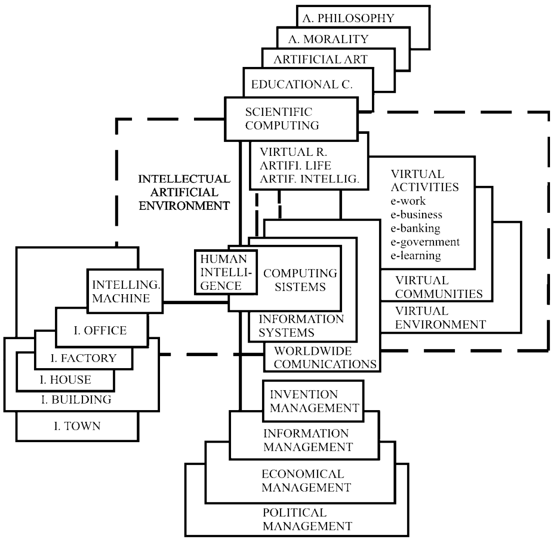


Fig.1 The artificial intellectual environment

1. Topical evolutions in cognition forms, techniques and models
Cognition forms are themselves evolving, in the scientific, philosophical, moral, aesthetical or political fields of culture. As an example, moral cognition and other components of moral culture, consciousness and spirituality were studied by us in other papers, regarding both artificial and human moral agents.

All these fields of culture are built by forms of cognition with some new traits in our days. These are, thus, synthetic and also evaluative forms of cognition. The term "synthetic cognition" means here not a result of an artificial construction, but a complex reunion, in a new whole, of different ways of cognition. We also can note the presence and action of an affective cognition, as a secondary result of the manifestation of affective intelligence and as a necessary component of beliefs. Beliefs, as cognitive, evaluative and affective complexes are, in their turn, aspects of consciousness, considered as a distinctive characteristic of humanity.

Inter-theoretical thinking

Meta-theoretical thinking

Systemic thinking

Integrative thinking

Probabilistic thinking

Fuzzy thinking

Heuristic thinking

Algorithmic thinking

Formal thinking

Artificial thinking

New forms of thinking are, among others: *meta-theoretical thinking, inter-theoretical thinking, integrative thinking, global thinking, projective thinking, information thinking, network thinking* and even *artificial thinking.*

We can add some forms of thinking already used but today developed in an unprecedented way, such as: *probabilistic thinking, heuristic thinking* and *algorithmic thinking,* implied in some of the above-mentioned new forms of thinking, under the influence of the usage, development and study of the information machine.

Some other forms or varieties of thinking, born by logical or mathematical invention, and used as scientific, artistic or philosophical methods and techniques, are *fuzzy thinking*, *fractal thinking* and *small-world thinking*.

It is also necessary to mention that what we call here *information thinking* is a *generic form and component of intellectual activity*, related especially to the information technologies. But we can also admit that all present or past forms of thinking are means of information processing.

Not all the afore-mentioned forms of thinking are generated by IT; some of them can be the result of some new evolutions in the mathematical or philosophical thinking, or manifestations of general kinds of scientific thinking, such as *systemic, probabilistic* or *formal thinking*.

Information thinking is determined by the development of information activities, the information machine and the information environment. Because these are now the most influential social factors and conditions, information thinking is the dominant form of thinking. We can even assert that *IT facilitates and even imposes a new conceptual thinking.*

This conceptual ensemble, described by more than the afore-mentioned ways of thinking moves along a few general trends, which determine privileged positions for some groups of values. The most general trends from all the observable ones are a) informatization of all aspects of human structures, relations and actions, b) intellectualization of various forms of activity, and c) essentialization of the intellectual activity. The above synthesized conceptual trends can be illustrated by some characteristics, forms, components, manifestations and representative values of the conceptual level of the intellectual activity.

The mentioned forms of human cognition need for their manifestation not just intellectual skills, but constitution and use of some psychical aptitudes, personality traits and cultural orientations and attitudes.
But spirituality, not just consciousness is needed for an effective artificial cognitive agent. This supplementary request for the conception and construction of artificial agents is necessary in order to allow common action of human and artificial beings in the present as well as in the future artificial cognitive environment.
In this new cognitive environment, new philosophical fields appear, such as the philosophy of computing, with its own domains, as well as the philosophy of computer science, philosophy of virtual reality, philosophy of artificial life, philosophy of communication mediated by computer, philosophy of the computer environment and others.
Scientific computing (which includes not just theorem demonstrating, but even chemical experiment techniques), is now continued in the philosophical field of research by Information Aesthetics, Artificial Morality or Digital Politics, all these opening the way towards the Artificial Philosophy.
This process of new philosophical field's appearance is associated with the constitution of new philosophical techniques in many important domains of reflection.

- Metaphysics is sometimes taken as *explanatory technique* for Prime Principles
- *observation*
- Epistemology studies *techniques of* *experimentation and*
- *verification*
-
- Formal axiology uses *value measurement techniques*
- Philosophy of language defines *language* as *technique of techniques*
- Social sciences are founding the *art of living* and even some
- *techniques of happiness.*
- Here work is presented as the main *technique of living*

Fig.3. New cognitive techniques in philosophy

Cognitive progresses are analysed at diverse levels by philosophical and scientific methods. The philosophical level which includes besides domains as General theory of knowledge, theory of scientific knowledge as well as philosophy of sciences even a philosophy of technology is continued by the scientific one which contains the so called "cognitive sciences" represented by branches of

psychology, of neurosciences, linguistics and others. This scientific approach is connected with research areas of computer science derived from the study of system's complexity and dynamics and developed by new domains as robotics, artificial intelligence and neural nets, machine vision and speech, machine learning, communicating and knowing.

From a philosophical perspective, even if not just theoretical but also empirical knowledge is recognized, and not only the abstract but often some concrete knowledge forms are admitted, the model of cognition is a rationalistic and spiritualist one. Interpretative and evaluative, prospective and prescriptive, anticipative and practical knowledge forms are recognized only if theoretically founded, scientifically deduced and practically verified. Sensation, perception, representation are not considered knowledge levels, because knowledge is identified with assertion of a true sentence.

In the area of cognitive psychology, in the contrary, multiple models of perceptive knowledge were built, such as the Grid model, the Prototype model, the Distinctive attributes model, the Holistic model, the Constructivist model or the Computational model.

Computer science offered the foundations and the means to conceive and to realize a perceptron and other noetic machines dedicated to form, voice, colour and even emotion detection and interpretation, to signs and meaning recognition, interpretation and use or to the construction of complex machines for special needs like "tactile vision" as well as to design super-sensorial robotic systems.

The nature as well as the genesis and the structure, the dynamics and the value of knowledge were already studied and synthesized in successive cognitive models, such as the 3C Model of knowledge; knowledge means, in this model, cognition, communication and co-operation.

Knowledge can also be explored through a model which conjoins the study of its generation with the analysis of its emergent internal structure and dynamic and with the prognosis of the generation practical, differential and total use of its products.

In all these aspects of the cognitive process the main role devolves upon cognitive techniques used in all domains of knowledge generation in knowledge work by knowledge groups and communities as well as in knowledge management and its domains and forms.

Our emergent Human-Artificial Knowledge Generation and Management model envisages in an interdisciplinary manner the collaborative cognitive processes which occur in the virtual technical intellectual environment of the computer culture.

The co-operative knowledge generation processes specific for our days suppose and include some mixed, human-artificial collaborative groups or communities made possible by knowledge based intelligent agents.

These processes are also conditioned by the present evolutions in the forms, levels and goals of the human cognitive processes, as well as by the arising of artificial knowledge and by the promising fusion between human and artificial knowledge. We will illustrate a few aspects of the three processes.

New knowledge is a result of knowledge work, knowledge networks and knowledge oriented techniques used in knowledge groups which are mastering a

new component of the computer world, a component born after hardware, software and middleware: the groupware.

2. Intellectual Techniques for Knowledge Work Knowledge Work and Groupware

Changes are occurring both in the elementary and upper levels of the intellectual activity.

- INFORMING ACTIVITIES are now developing by
  - integral representation of the "map" of a net of knowledge
  - reconstitution of a line of search by recording the successive links
  - use of "reading marks"
  - means to edit, to multiply and transmit scientific information
- RESEARCH ACTIVITIES are improved by
  - Text mining instruments, Knowledge Discovery in KB
  - Collocation detectors, Syntactic or semantic annotators, summarizers
  - Semantic search and classifiers
  - Specialized semantic editors by search engines and ontologies; ex.: the *GenWeb* System
- FORMATIVE ACTIVITIES use
  - Narrative learning environments on the web,
  - Immersive contexts for learning foreign terminology
  - Personalized knowledge –based e-learning environments
  - Instruments for knowledge extraction from the web

- CREATIVE ACTIVITIES are now assisted by computer work and skills which develop
  - scientific and technical logic
  - information aesthetics
  - computer ethics
  - philosophy of digital politics
  - philosophy of ICT.

  } artificial philosophy

Implied in all these types of activities, knowledge work needs some specific intellectual techniques.

- INFORMING TECHNIQUES
  - Lecture assisted by computer
  - Hypertext as a reading and writing technique; presupposes other techniques

- KNOWLEDGE DISCOVERY TECHNIQUES
  - knowledge extracting techniques from the web
  - co-constructed narratives in scientific problems solving
- FORMATIVE TECHNIQUES
  - Knowledge transfer techniques
    - format change
    - storage media change
    - translation techniques: computer assisted, automatic, author assisted
  - Organizational learning techniques assisted by
    - intelligent tutoring systems
    - knowledge based, intelligent and flexible systems
      ex.: the *SINTEC Personalized , Knowledge-Based E - Learning Environment*

- KNOWLEDGE TRADE AND KNOWLEDGE MANAGEMENT TECHNIQUES

Knowledge management has as object the result of intellectual activity; KM techniques are Intellectual Techniques

Diverse forms of knowledge work such as

Collaborative knowledge generation
Knowledge representation
Knowledge discovery in KB and Knowledge Warehouses
Knowledge Processing & Knowledge Transfer and even

Knowledge distribution
are finalized by knowledge management and its forms.
Between the aims of knowledge management are the
• integral use of knowledge, even of

- implicit
- ignored knowledge
- lost

• allocation of the totality of knowledge between different levels and forms of social action structures

Knowledge work is / can be
Virtual work
Work at distance
Group work
Human
Artificial
Mixed

Knowledge Management

Cognitive techniques are now multiplied at each level of the intellectual activity, from the informing to the creative activities. Searching techniques are used by webbots on the Net, learning techniques are developed for artificial tutoring agents, other intellectual techniques are conceived and implemented for research, design, management or banking activities, for e-business and e-commerce, for e-anything.

By common activities in this new cognitive environment, human and artificial cognitive agents are gaining new common traits. Both types of cognitive agents are or will be:

1) individual entities (complex, specialized, autonomous or self-determined, even unpredictable ones), 2) open and even free conduct performing systems (with specific, flexible and heuristic mechanisms and procedures of decision), 3) cultural beings: the free conduct gives cultural value to the action of a human (natural) or artificial being, 4) systems open to education, not just to instruction, 5) entities with "lifegraphy", not just "stategraphy", 6) entities endowed with diverse or even multiple cognitive skills and techniques, 7) equipped not just with automatisms and intelligence, but with beliefs (cognitive, evaluative and affective complexes), 8) capable even of reflection (cultural life is a form of spiritual, not just of conscious activity), 9) components/members of some real (corporal or virtual) community.

Classes of attributes which are referring to a) sensing and acting, b) reasoning, c) learning and knowing, d) their structure and e) number, are distinguished for intelligent agents by specialists in theory and methodology of Intelligent agents [Skolnicki and Arciszewski, 2003].

About individual agents these authors mention that they act locally, cooperate, are sophisticated, they do not model other agents and do not show internal state, are trustful and acquire knowledge, have stable architecture and work in group. Swarm agents are more locally, share resources, have less autonomy but are more competitive and more mobile, react more directly and may discover roles in runtime, use fixed language and assume information to be true, are less transparent and less reusable.

Human and artificial agents can be compared as intelligent agents.

	Intelligent Agents	

Human		Artificial
• initiative • subject and object of action • social • reflective • omni-oriented • in(de)finitely perfectible		• reactive • pro-active • sociable • self-analytic(al) • guidable • teachable

Fig 4 Features of intelligent agents

An intelligent agent, human or artificial, is characterized by its actions, by its environment and by the events at which it participates or which it generates, by its beliefs, plans, goals and by its levels and forms of communication. In our artificial intellectual environment both human and artificial intelligent agents compose and transmit messages observing some protocols.

Intelligent agents as learning and knowing agents are distinguished by their goals and by some features which ensure their capacity to accomplish these goals, such as adaptability, specialized structures for knowledge storing, ability to establish the true value of knowledge, as well as by their capacity to use knowledge: to learn on the acquired knowledge base.

Knowledge can be used both for their own evolution and for specific problems solving. As an example for the second case, we can show the scheme created by some scientists to explain the role of AI using in virtually all kinds of Business.

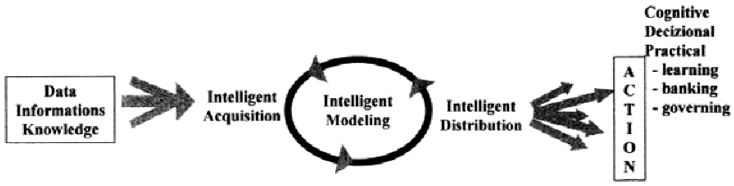


Fig.5. A framework for intelligent agents based business intelligence

New cognitive techniques are used in each field of the present artificial cognitive environment. Some of these techniques can be common for artificial and human cognitive agents.

3. Common activities, tools and environments of human and artificial cognitive agents

Multiplication, intrication and evolution of knowledge forms are between the factors which induce, in our days, a common evolution of human and artificial intelligent agents.
Human knowledge evolves in our days by many and often opposite forms; their opposition is a permanent source of new ways of cognition generation and use. Human intelligence is assisted by artificial intelligence forms and, consequently, by artificial knowledge. More, we are faced with a new knowledge level - the process and the result of the common knowledge work of human ad artificial cognitive agents: the human-artificial knowledge.

Human knowledge	Human and artificial knowledge	Human – artificial knowledge
- Systemic knowledge – Fuzzy knowledge - Global knowledge – Fractal knowledge - Network knowledge – Local knowledge - Projective and Heuristic knowledge - Intra-theoretical and Integrative Knowledge - Knowledge networks → New knowledge	To KNOW ON THE WEB • using knowbots • co-operating with Intelligent Agents ◊ E-LEARNING with Knowledge Based Intelligent Agents ◊ To make KNOWLEDGE MANAGEMENT by "business intelligence"	• Knowledge representation • Knowledge structuring: 'knowledge making" "knowledge management" • Knowledge generation • Knowledge dissemination • Integration of different knowledge techniques by: - multi-agent systems - communication facilities on ɔal hypertext of the web

Fig. 6 Aspects in the evolution of forms and possibilities of knowledge

Evolution can be development or involution. What will it be the co-evolution of human and artificial cognitive agents: development or involution? Today human intellectual abilities in their individual and group manifestations are evolving both towards excellence and towards the lack of performance.
The past *natural evolution* of humans lasted 2 or 4 000 000 years, and their *cultural evolution* - 8 000, 50 000 or maybe 350 000 years. An evident acceleration of the cultural cycles was recorded along these uncertain but long intervals. This acceleration means not only the shortening of the historical

period's length, but the growth of the development rhythm in the same time. Both the natural and cultural evolution seem to be spontaneous evolution forms. *Technological evolution* is considered, on the contrary, as a directed evolution. According to some computer scientists, evolution of engineered systems in general occurs conforming to patterns of evolutions, is a directed evolution or, at least, understanding of present stages of evolution can be a step to predict further evolution and even to "speed up" the process.

Current stage of IAS is established by computer scientists (see Skolnicki and Arciszewski) along with specific criteria, such as

- run-time acquisition of knowledge
- growing number of features
- growing flexibility and controllability
- starting simplification
- internal architecture
- general use
- decreasing human involvement and automation.

Theories about CULTURAL EVOLUTION of Artificial (Cognitive) Agents:

- Evolution by SIMULATION of NATURAL EVOLUTION of populations including
  - ecosystems
  - mutations
  - viruses
  - selection
- Self – structuring in ordered context: engineers will create just suitable conditions
- Learning by cultural processes like children

Stages in the evolution of Artificial Agents are steps in Human Evolution, state some technologists.

by

HUMANS evolve ARTIFICIAL (COGNITIVE) AGENTS

with

Past and Present Models of Artificial Cognitive Agents Development	• BRAIN as a model • MIND as a model • KNOWLEDGE as a model

If HUMAN and ARTIFICIAL Cognitive Agents are CO-EVOLVING are these models adequate for a common evolution? Can separately an accepted model of human cognitive agents development and a model of artificial cognitive agents development explain and foresee their real co-evolution?

Both the signaled co-evolution and its explanation, interpretation and control need a new theoretical model, a model founded on a vision about a new group, community or even world sized cognitive agent, with natural and artificial, biotic and technical, material and virtual, individual and social components.

The holistic knowledge theories stated in the last century that the unity of knowledge is the whole knowledge, that the concept of truth is defined according to a conception about the truth and even that each truth has sense between the limits of a theory of the truth.

We can say, by similitude, that *by the analysed convergent cognitive processes, a new cognitive agent was born*, and that this new cognitive unity realizes, in the same time, the most profound unification of the natural and artificial by a new perspective about the ideal, virtual and spiritual.

Directed or just foreseen, co-evolution of human and artificial (cognitive) agents as well as their possible /desirable convergence presupposes other cognitive and cultural processes:

<table>
<tr><td>KNOWLEDGE DIVIESIFICATION:
• social knowledge
• prospective knowledge
• self-cognition
CULTURE LEARNING: values
◊ understanding
◊ sharing
◊ practicing
CREATION OF VALUES:
o Practical (technical, moral, political)
o Intellectual(scientific, philosophical)
o Spiritual (artistic, religious)
INVENTION of
 Behaviour
 Institutions</td><td>SPIRIT as a model?</td></tr>
</table>

Ideas Self Societies Future	

The co-evolution of human and artificial cognitive agents occurs by many different but connected ways: structural and functional, spontaneous and directed, abstract and concrete, material and ideal, cognitive and practical ones.
By cognitive techniques used in a new technical intellectual environment human and artificial agents are gaining some common features.
Cognition forms are themselves evolving now in all fields of culture and even new forms of knowledge appear.
Some common cognitive techniques are used by human and artificial agents which are forming new communities by knowledge work in knowledge groups and knowledge networks.
Applied by humans and machines, who can meet in the middle of the road between the natural and the artificial, these cognitive techniques can facilitate a common, faster evolution.

References

Bedau, M. A. (1998) „Philosophical Content and Method of Artificial Life", in Bynum, T. W. and Moor, J. H. (eds.), *The Digital Phoenix: How Computers are Changing Philosophy*, Blackwell Publishers, Oxford.

Bynum, T. W. (1998) „Global Information Ethics and the Information Revolution", in Bynum, T. W. and Moor, J. H. (eds.), *The Digital Phoenix: How Computers are Changing Philosophy*, Blackwell Publishers, Oxford.

P. M. Churchland (1998) „The Neural Representation of the Social World", in Bynum, T. W. and Moor, J. H. (eds.), *The Digital Phoenix: How Computers are Changing Philosophy*, Blackwell Publishers, Oxford

Danielson, P. (1998) "How Computers Extend Artificial Morality", in Bynum, T. W. and Moor, J. H. (eds.), *The Digital Phoenix: How Computers are Changing Philosophy*, Blackwell Publishers, Oxford.

Ey, H. (1982) *Conştiinţa* (The Consciousness), Editura Ştiinţifică, Bucureşti

Floridi, L. (1999), „Information Ethics: On the Theoretical Foundation of Computer Ethics", *Ethics and Information Technology,* 1.1, p. 37-56.

Gregory, R. (2000) *Viitorul creatorilor de inteligenţă* (The Future of Mind-Makers), Editura Ştiinţifică, Bucureşti

Hillis, W. D., *Maşina care gândeşte* (The Pattern on the Stone. The Simple Ideas that Make Computers Work), Editura Humanitas, Bucureşti, 2001

Laruelle, Fr. (1990) *Théorie des identitées, fractalité generalisée et philosophie artificielle,* P. U. F., Paris.

Pană, L. (2005) "Filosofia artificialului şi filosofia artificială" (The Philosophy of the Artificial and the Artificial Philosophy), in *Academica*, nr. 34, ianuarie, Anul XV, 171.

Pană, L. (2004 b) “Etica artificială" (Artificial Ethics) in *Filosofia informației şi a tehnicii informaționale* (The Philosophy of Information and Information Technology), Editura Politehnica Press, Bucureşti

Pană, L. (2006) “Knowledge Management and Intellectual Techniques. Intellectual Invention and Its Forms”, in R. Trappl, (ed.), C*ybernetics and systems,* volume 1, *Proceedings of the Eighteenth European Meeting on Cybernetic and Systems Research,* organized by the Austrian Society for Cybernetic Studies, held at the University of Vienna, Austria, 18-21 April 2006.

Pană, L. 2005 “Moral Intelligence for Artificial and Human Agents”, *Machine Ethics, Papers from the AAAI Fall Symposium Series*, Arlington, Virginia, 2005, November 4-6, AAAI Press, Menlo Park, California.

Pană, L. (forthcoming) “The Intelligent Environment as an Answer to Complexity”, *Proceedings of the XV IUAES Congress “*Humankind/Nature Interaction: Past, Present and Future”, Florence, 2003, in the volume *The Trans-disciplinary Flow of Our World*

Sarmiento, J., Trausan-Matu, St., Stahl, G., (2005) *Co-constructed Narratives in online, collaborative mathematic Problems Solving,* in *Proceedings of Narrative Learning Environments Workshop,* 12th International Conference on AI IN Education (AI – ED 2005), Amsterdam, Netherlands, p. 56-62, 2005.

Skolnicki, Z. and Arciszewski, T. (2003) "Intelligent Agents in Design", in *The 15th International Conference on Design Theory and Methodology*, George Mason University, 2-6 September 2003, Chicago, Illinois

Sloman, A. (1990) “Motives, Mechanisms, Emotions”, in M. Boden (ed.), *The Philosophy of Artificial Intelligence,* Oxford University Press.

Sowa, J. (1999) *Knowledge Representation: Logical, Philosophical and Computational Foundations*, Brooke Cole Publishing Co., Pacific Grove, CA, 1999.

Trăuşan-Matu, St., Maraschi, D., Cerri, S., Ontology-Centered Personalized Presentation of Knowledge Extracted From the Web, in S.Cerri, G.Gouarderes (eds.), Intelligent Tutoring Systems 2002, Springer, Lecture Notes in Computer Science number 2363, ISBN 3-540-43750-9, pp.259-269.

Trăuşan-Matu, St., Immersive Contexts for Learning Foreign Terminology on the Web, *Proceedings of WITREC-2000*, Montpelier, France, http://www.lirmm.fr/WITREC

Trăuşan-Matu, St., (2000) *Metaphor Processing for Learning Terminology on the Web*, in S. A. Cerri (ed.), Artificial Intelligence, Methodology, Systems, Applications 2000, Springer-Verlag, ISBN 3-540-41044-9.

Trăusan-Matu, St., Cristea, V., Udrea, O., (2005) *The SINTEC Personalized, Knowledge-Based E-Learning Environment,* in C. Niculescu, A. Ionita (eds), International Simposium on Organizational Learning and Knowledge Work Management (OL – KWM 2005), Bucharest, Romania, ISBN 973-618-058-1.

MODELLING INTUITION AND INTENTIONALITY FROM A NEUROPSYCHOLOGICAL PERSPECTIVE

Dipl.-Psych. Tino Schmidt
Technical University Dresden
Dept. of Psychology
Zellescher Weg 17
D-01069 Dresden, Germany
Email: schmidtt@psychologie.tu-dresden.de
Tel.:+49-351-4633-3824

Dipl.-Psych. Hagen Lehmann
University of Bath
Dept. of Comp. Science
Bath BA2 7AY U.K.
Email: H.Lehmann@cs.bath.ac.uk
Tel.:+49-1225-386183

The concept of intentionality is strongly debated in the field of computer science. There are different points of view, but overall it is assumed, that intentionality is a component of adaptive intelligent systems.

In the field of neuroscience the debate on the same topic is less strong. This is probably due to the difficulties doing research on intentions, since they are only accessible from a first person perspective as a "feeling of control" in our everyday life.

We present a model which shows a way to make intentionality accessible from a third person perspective. It is based on results of several neuropsychological studies and will enable us to make the phenomenon of the "feeling of control" measurable.

In order to proof the model empirically we conduct a series of experiments with an EEG based "brain-computer-interface" with which we are able to influence the "feeling of control".

Keywords: intentionality, brain-computer-interface, feeling of control, effect anticipation, feedback loop, consciousness

Theoretical background

1.1. Definition of the “sense of agency“

How is it possible to be certain, that we are the source of any of our voluntary actions? In everyday life we are able to interact with our environment without asking this question. Our cognitive system gives us an immediate sense of authorship for our own actions and the ability to differentiate them from actions done by other agents in the environment.
This ability is often described as “sense of agency”. (de Vignemont & Fourneret, 2004)
We think that this ability is one of the key components towards a better understanding of the underlying evolutionary mechanisms of the development of our cognitive system and might therefore help us in the development of artificial cognitive systems. (Blakemore & Decety, 2001)
We assume that the central nervous system contains internal models, which represent aspects of one’s own body and its interaction with the external world in order to optimize motor control and learning.
Our model focuses on two types of internal models. The ‘forward loop’ uses efference copies to predict the sensory consequences of motor commands whenever movements are made. The ‘feedback loop’ provides the current action program with corrections of the motor commands to achieve the state desired by the action.
In this paper, we put forward that the coordination between the feedback-loop and “feedforward”-loop forms the “sense of agency”. We propose a model of how this coordination could be realised in artificial intelligence, based on the functional principles of the human brain. To illustrate our model, we will show some neurophysiological evidence for the phenomena to be discussed.

1.2. Is the “sense of agency” a key component of higher cognitive functions?

The ability to understand ourselves as the cause of an action enables us to attribute effects caused by this action to ourselves. One therefore experiences the own self not only as the cause but also as the result of the same action. (Wegner & Wheatley, 1999) This seems to be a small difference, but it has an enormous effect on the possibilities of the organism to manipulate its environment instead of only reacting to changes in it. It gives the organism the ability to abstract rules from its

surroundings, act upon them and modify its goals according to possible changes.

The basic property of a nervous system is the distinction between different states of the environment or of the own body at different points in time. With the ability to differentiate between actions done by oneself and environmental effects, the organism has the preconditions for the development of the “theory of mind” and with that a “sense of agency”.

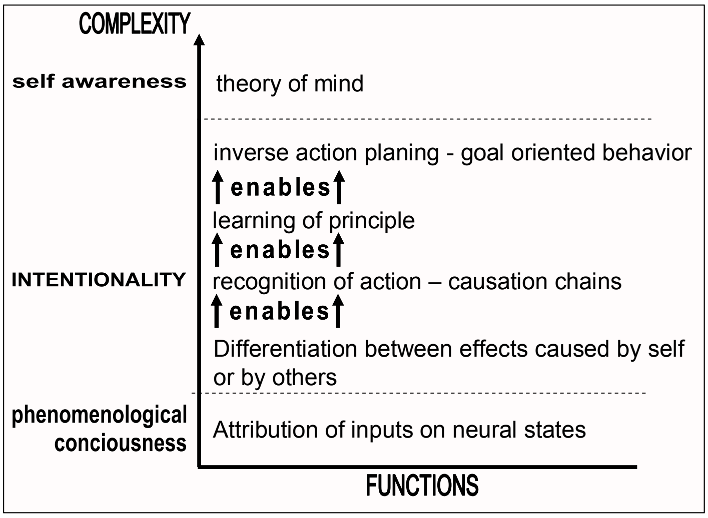


Fig.1 Taxonomy of fundtions of consciousness

Thus, the “sense of agency” is the connection between the basic function of the nervous system and the complex patterns of social behaviour and cognition.
We, therefore, assume that the “sense of agency” is more than only a philosophical construct used in the discussion on artificial intelligence, but that it is a likely embodiment of the central precondition for the creation of artificial systems with higher cognitive functions.

There must be an environment in which it is possible for the self to differentiate itself from its surroundings in order to understand itself as an acting entity in that environment.

The difference between an environment and an intelligent system is that it acts towards a certain goal. This is what we shall call the "intentionality of an artificial cognitive system".

2. The static model

2.1. Subject of the static model

Our model is based on the model by Wolpert (Wolpert et al, 1995), which describes how actions are initiated and controlled by the human brain.
In his model, he proposes that an action is executed by an action program and is initiated to achieve a goal.
The central question of his work is how motor action is initiated and controlled in interaction with the external world in order to optimise the process of action selection in relation to the current goals of the organism. (Wolpert & Ghahramani, 2000)
He analyses two internal models of the process of motor control. The "internal forward model" is a model within the brain that can predict the sensory consequence of an action. The 'inverse model' provides the motor commands with the processed feedback about the actual state of goal-achievement.
Wolpert views motor learning as the adaptation of forward and inverse internal models appropriate for different tasks and environments. (Wolpert & Kawato 1998)

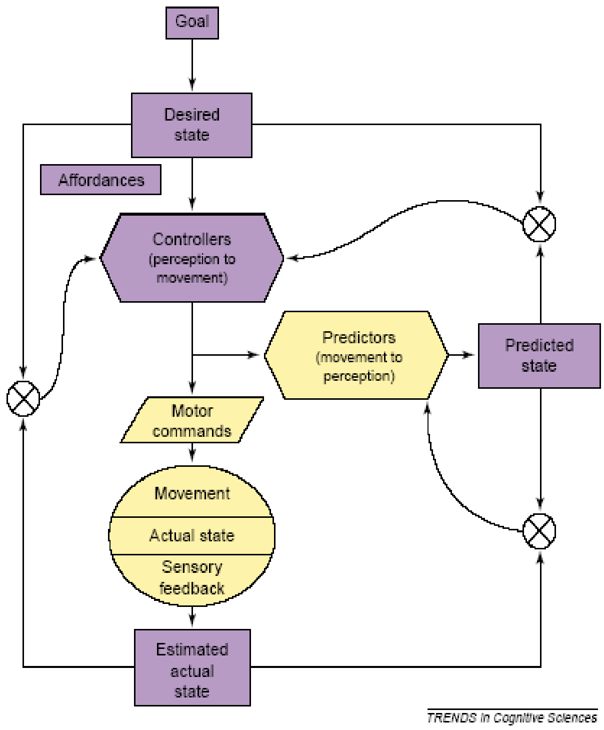
2.2. Description of the static model

The forward model predicts the actual outcome of motor commands and compares it to the desired outcome. The desired outcome describes a specific current goal of an organism and is called the "desired state". The prediction of the hypothetical result of the current action is called the "predicted state". The comparison between the "predicted state" and "desired state" occurs before the movement is initiated.
The prediction itself is used to estimate the state of the motor system to make fine adjustments before reafferent feedback from the movement is

available. The point of reference of the reafferent feedback is called "actual state". In fact, the "actual state" represents the percepted effect of the action. Thus, the comparison between the "actual state" and the "desired state" represents the inverse model while the forward model predicts the sensory consequences of movement and compares this with the actual feedback. (Wolpert, 1995)
This comparison occurs after a movement is made. The resulting prediction of the effects of an action can be used to anticipate and compensate for the sensory effects of movement. The planning (specification of the trajectory) and execution of a movement are parallel running dynamic processes.

Thus, the central result of Wolperts idea is that the estimated position of a limb is not solely based on sensory information within the "inverse model", but also on the stream of motor commands issued to the limbs. (Wolpert et al, 2001)
On the basis of these commands, the forward model can estimate the new position of the limb before any sensory feedback has been received. The experience of moving the limb is based on the comparisons between the "predicted state" as the central point of reference in the model and the both other states.

Fig. 2: Wolpert's model



2.3. Critique of the static model

Wolpert's idea is that forward models can be used to provide an optimal estimation of the position of the body in the environment, and might even be used for mental practice. Forward models can be used to explain why we are unable to tickle ourselves. (Blakemore et al, 2001) This characteristic is our starting point in explaining agency in the context of motor control.

The main problem of Wolpert's model is how to explain agency. There is no definition of the circumstances under which the components of internal models gain the ability to rise into consciousness. (Frith et al, 2000) There is also no clear specification of how the comparisons between desired state, predicted state and actual state are coordinated in time. We refer to Wolpert's model as a "static model" because he describes a dynamic process without describing how the dynamics work in time (Spence, 1996). Consequently, there is no clear definition of how the comparisons, as the dynamic components, enable a mechanism to identify whether an effect comes from its own action in the environment or is internally generated.
The aim of our model is to explain how to identify the effects of our own actions. Thus, we will elicit the dynamics of Wolpert's model.

3. The dynamic model

3.1. Components

We propose that each feeling is the result of a hierarchy of several in nested loops of neural pattern matching. Each of these processes consists of a "feedforward"-loop, feedback – loop system, matching neural pattern codes of a former time period with current patterns. (Jeannerod et al 1995).

3.1.1. Feedback loop
We assume that the initiation of a motor action follows a continuous feedback-loop with the environment. The task of this feedback-loop is to fit actions on situations in a permanently changing environment. The central element of this process is action selection.

During the action selection process particular action programs are chosen, which lead, in their execution, to the predictive states, which correspond best to the desired states.

The question arises of how this process works. Our starting point is a natural nervous system from which we assume that each activation pattern incorporates in the entirety of all neurons a function of the behaviour and experience of the organism.

The desired state as well as the predictive state is both neuronal states and, therefore, activation patterns of the nervous system. The better the resemblance of these two activation patterns, the more likely is the activation of the according action program. (Jeannerod, 1999)

Along with the activation of this action program, the pattern of the predicted state of the action is copied as „temporal actual state“.
This process has two effects:
1. It is not possible to tell whether the activation pattern follows perception, i.e. the actual state, or is self-induced, i.e. the predictive state. If the active state were equal to the predictive state, the action would proceed automatically. This process would have the ability to rise into consciousness on a higher cognitive level, but not necessarily so. The process of breathing, for example, is in most cases unconscious still it is easy for us to become aware that we draw breath.
2. The action selection mechanism starts all over again if no matching can be found between both patterns. An alternative action with a supposedly better matching of the patterns of the predicted state and the changed temporal actual state is chosen. The consequence of such a change in the action selection process would enter awareness.
A sudden interruption of our breathing would be an example for such a case.
The interesting point is, that the action program started unconsciously, but we know at the moment of its interruption that we are the cause of what happened.

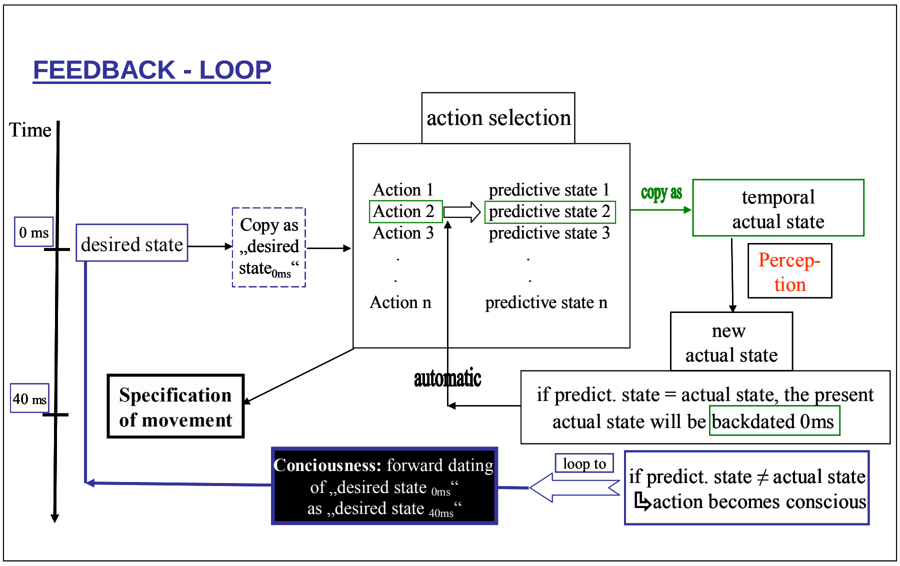


Fig. 3: Feedback – Loop of the dynamic model

In order to become aware of an automatized action, in the aftermath of this action, the initiation of the automatized action and the perception of its result have to be experienced simultaneously.
The question is of how it is possible to perceive us as the cause for an action program, if the point of its initiation and accordingly the point of reference for the feedback-loop happen unconsciously.

3.1.2. “Feedforward” loop

To solve this problem, we assume that this process is not a feedback-loop but a “feedforward”-loop. “Feedforward” means that the predictive state of the chosen action is coded already at the point of the initiation of the action as „temporary actual state“.
This means, on a physiological level, that the activation pattern which is supposed to follow the action (predictive state) is activated simultaneously with the action itself. It also has to be stored for as long as the action is executed. (Georgieff & Jeannerod 1998)

Thus, the hypothetical result of an action is anticipated via the process of action selection.

If the anticipated neuronal result and the following action pattern of the „actual state“ correspond with each other, it is possible to psycho-physiologically backdate this perception as “original predictive state“.

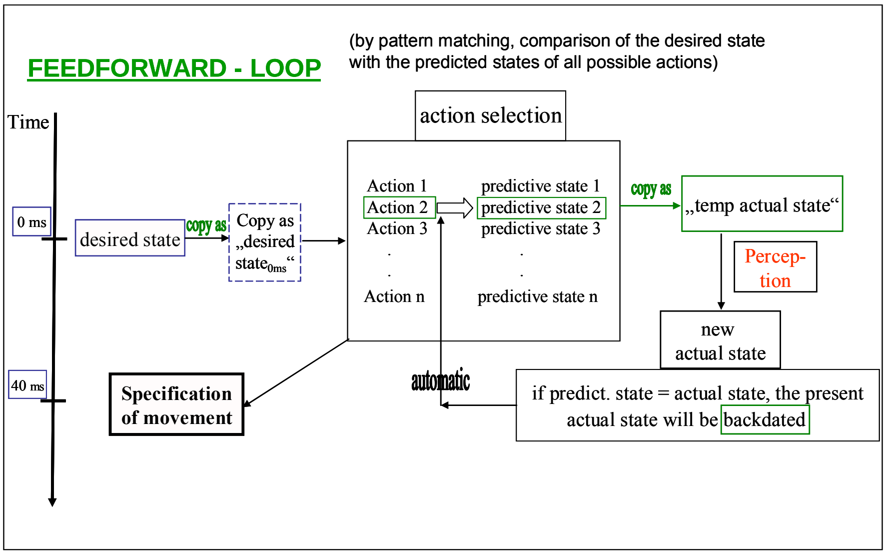


Fig. 4: Feedforward – Loop of the dynamic model

Consequently, our psychological time should differ from the actual physical time. As long as our perception of time is coherent we do not recognize the difference. We perceive a point in time as contemporary when it is already in the past.
We become aware of this discrepancy only in a few situations. If we touch a hot oven plate, for example, reflex action makes us draw back our hand, but we feel the pain only at the moment after the reaction. In this case, the reaction owing to the pain happens before the perception of it.
This is a side effect of the “feedforward”-loop. The capacity for fast reactions, an advantage of automatized actions, is maintained. But what is the advantage of the delayed awareness as effect of the “feedforward”-loop?

We propose that the benefit of the “feedforward”-loop for the organism is the maintenance of both fast reaction times and the adaptability of conscious actions in a dynamically changing environment.

3.2. Coordination of the components

Two effects follow if the predicted state of the selected action does not correspond with the delayed perceived actual temporal state.

1. An action selection process has to be initiated to adapt the organism to the changed environment with the help of a new action selection program.
2. The organism has to perceive itself as the source of that change; otherwise it would feel controlled by an alien force.

This process is connected with the awareness for the changes in the hitherto automatized actions. But a change in the states of the actions can only become aware if memory for past states exists. During an automatized action, this problem was solved with the backdating of the "temporal actual state". This was possible because the "actual state" was, in this case, identical with the permanently activated "predictive state".
In our case this is not possible, because the change of action is caused by the difference between the pattern of the predictive state and the actual state.

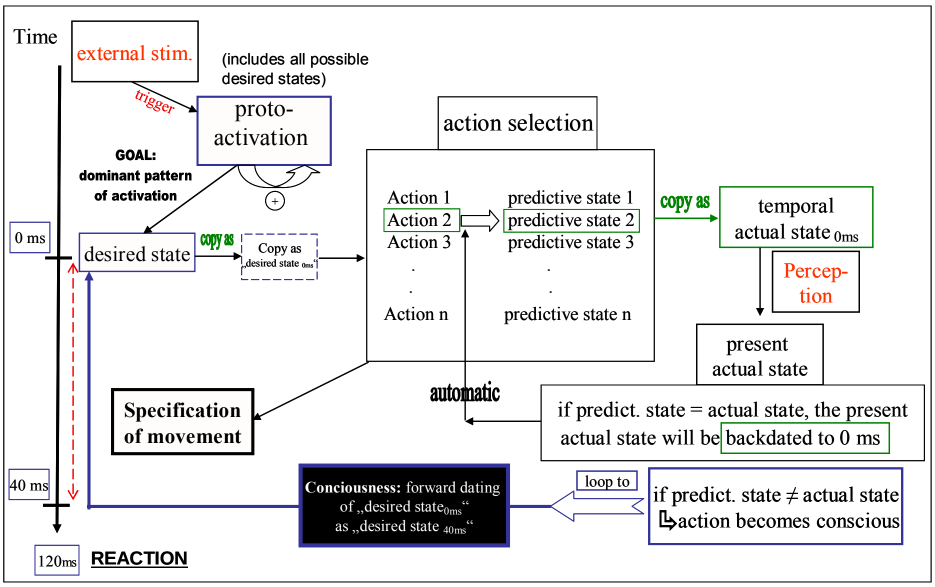


Fig. 5: Flowchart of the dynamic model

How is it possible to maintain the „sense of agency"?

We think that the desired state, which initiated the original, automatized action is dated forward after the action happened. In the process of becoming aware of an action, the neuronal activation pattern underlying the action is addressed in time.

The discrepancy between the perceived actual state and the predictive state of an action is recognized in the prefrontal lobe as specific activation pattern. This activation pattern itself reinforces the activation pattern of the original desired state in the prefrontal lobe. This mechanism is theoretically conceptualised as a top down process. (Jeannerod, 2003)

We become aware, not of the action but the goal of this action. The action itself can be modified according to the goal of this process. During the following new action selection process, no complete action programs are activated, but the actual action program is modified in specific points. The organism learns while acting. Complete action programs are continuously modified in order to adapt the organism to an environment with permanently changing conditions.

We do not need to assume that every perceived state of the environment is represented by exactly one neuronal pattern.
We do assume that the permanent system-innate activity is based on prototypical activation patterns of the entire network. These activation patterns consist of every prototypical action and perception pattern of the organism. Prototypical here means inherent general samples like the shape of edges.
An example from neurophysiology are the Hubel-Wiesel-Receptor cells. They are activated during the perception of any kind of edge.
In our view, it is not the cells that are the neuronal correlate of the perception but the change in the activation pattern of the entire neuronal network, which is triggered by the activation of these cells.

In this context, learning means the specification of the prototypical system activations. Learning is the adaptation of given action programs to different environmental affordances.

This information processing structure enables the organism to interact fast and flexibly with its surroundings.

The model describes how the automatized actions are initiated, at what point in time we become aware of these actions and when they are modified.
It is not yet explained why this process of flexible action pattern modification requires consciousness and what function the „sense of agency„ has. (van den Bos & Jeannerod, 2002)

With regard to the execution of automatized actions, it is not important how the action is controlled; who or what is initiating the action. Automatized actions do not need a feedback of their effects to regulate themselves. They work within a constant frame, which consists of a stable action program. It is even contra-productive for automatized processes when the effects of automatized actions come into awareness. Professional dancers perform worse when they are forced to concentrate on their next steps while dancing. Jugglers who try to observe their juggling balls while performing are another example.

When adapting an action program to altered environmental conditions, we need to have a constant feedback from the environment. Furthermore, every alteration of an action program is based on the original automatized action program. This results in a time-based interaction between the feedback and the "feedforward"-loop. This interaction enables us to experience the sense of agency. We become aware of being the initiators of our own actions. If there is a distortion between the timing of both loops, our sense of agency is disturbed, too. (Frith et al 2000)

3.3. Empirical evidence

This can be illustrated by sample cases of psychological disorders:

1. Anarchic Hand Syndrome (AHS)

Patients with Anarchic Hand Syndrome are unable to control the actions of one of their own hands. In contrast to other forms of involuntary limb movements, the movements of AHS patients remain purposeful and goal-based. Some patients "personify" their hand and dissociate themselves from the behaviour of it by giving it a name and attributing a separate and distinct personality to it. The patients are aware of their behaviour but cannot inhibit it. In this case, there is an impairment of the temporal coordination between the feedback processing of both hemispheres of

the brain (Sala *et al*, 1991). As a result, movements of the anarchic limb do not fit with the patient's current goals.

2. Utilization behaviour (UB)

Patients with UB give stereotyped "object-appropriate" motor responses to environmental cues and objects, but mostly inappropriate ones for the particular context. (Eslinger *et al,* 1991) They take glasses, which they see, while wearing glasses at the same time - to give an example. But there is no awareness of goals and intended actions.

The actions are involuntarily elicited by objects in the environment, but the patient experiences these actions as intended, without any perception of the discrepancy between actions and intentions. The patients claim to intend the action.

In this case, automatized movements arise in the same way as top-down regulated movements. The result of this mechanism is an attribution of these movements to the own will, although the movements contradict their current goals.

UB results from a temporal imbalance between a proposed voluntary goal-directed, and future-directed feedback-loop motor system and an automatic, stimulus-bound "feedforward"-system.

The „voluntary control" of the feedback-loop is transferred to a fast, autonomous stimulus reaction. This mechanism forms the „illusion of control": the automatic reaction was freely chosen by the person.

3. Delusion of control

The most common case of loosing the sense of agency might be Schizophrenia, especially the syndrome of the delusion of control (Mellors, 1970).

The patient is aware of his goal, of his intention to move and of his current movements having occurred, but is not aware of having initiated this. Schizophrenic patients often feel controlled by an external agent. Sometimes the perception also underlies a delusion of control, and the patients are aware of hearing alien voices controlling their thoughts and actions.

While normally the sensory consequences of self-generated movements are classified as self-produced, in the case of delusion of control, the patient is not aware of the predicted consequences of a movement and is therefore not aware of having initiated it. They cannot clearly differentiate between their "inner voice" and the rehearsal of voices they remember. (Campbell, 1999)

3.3.1. Experimental results

In their study, Sato & Yasuda (2005) showed that the congruency between an action and its auditory consequence induced independently the experience of agency. In their first experiment, the sense of self-agency was reduced when the presentation of a tone was unpredictable in terms of timing and its frequency, although in fact the tone was self-produced. In the second experiment, the opposite was found. That is, participants experienced illusionary sense of self-agency when the externally generated sensations happened to match the prediction made by them. In the third experiment, the sense of self-agency was reduced when there was a discrepancy between the predicted and actual sensory consequences, regardless of presence or absence of a discrepancy between the intended and actual consequences of actions.

Haggard *et al* (2002) have demonstrated that an action and its effect are perceived as being closer in time when the consequence is intended. He used the perceived time of intentional actions and of their sensory consequences as means to study consciousness of action. These perceived times were attracted together in conscious awareness, so that subjects perceived voluntary movements as occurring later and their sensory consequences as occurring earlier than they actually did. Comparable involuntary movements caused by magnetic brain stimulation reversed this attraction effect. He concluded that the Central Nervous System applies a specific neural mechanism to produce intentional binding of actions and their effects in conscious awareness.

4. Discussion and perspectives

There is no problem to explain the effect of delusion of control by our model. The only problem might be, to explain why this discrepancy is attributed to an external agent?

It is possible, with our model, to explain the effect of delusion of control, even though it is difficult to understand why the discrepancy causing the problem is attributed to an external agent? (Daprati, 1997)

The psychological disorders described above illustrate different types of problems with experience of the sense of agency.

In the “Anarchic Hand Syndrome” the movements of the „bad“ hand arise automatically, but are not attributed to the own person (Sala et al, 1991), while in the “Utilisation Behaviour”, even though the movements arise also automatically, they are attributed to the own person (Lhermitte et al, 1986).

The source of authorship is the intention, which causes the action. The crucial point is the attribution of the intention to an entity. (Gerrans, 2002) The ability to attribute an action to oneself needs requires a concept of self.

Inverse action planning cannot be the reason for the evolution of the concept of self, because this would generate a circular argument. This is due to the fact, that the ability to differentiate between effects caused by oneself or by others is the precondition of inverse action planning.

The key to understand the evolution of the self could be the neural mechanism of imitation, because the understanding of goal directed action requires the ability to recognize which action is appropriate for which object in which circumstances as well as the ability to understand the intention of goal directed actions performed by others.

Rizzolatti and colleagues (1988, 1996) discovered two populations of neurones that discharge during goal directed actions (Rizzolatti et al, 1988). Among these neurones though they are indistinguishable as far as their motor properties are concerned.
The first type, called “canonical neurones,” are activated when the organism sees an object, as well as when it grasps that object. On the basis of these common attributes, it has been proposed that these neurons are involved in the transformation of intrinsic properties of objects into the appropriate hand movements (Jeannerod et al, 1995).
The second type, called “mirror neurones”, fire when an agent grasping an object is observed and when grasping the object itself.
To understand agency, further research should connect both perspectives: the interaction of the temporal dynamic of the interaction of the feedback- and “feedforward”-loops and the neuronal basis of imitation learning.

5. Bibliography

- Blakemore S.J., Wolpert D.M. & Frith C.D., 2000. Why can’t we tickle ourselves?, *NeuroReport* **11**, pp. 11–16.

- Blakemore S.J. & Decety, 2001. From the perception of action to the understanding of intention, *Nature Reviews Neuroscience* **2**, pp. 561–567.
- Campbell J., 1999. Schizophrenia, the space of reasons and thinking as a motor process, *The Monist* **82**, pp. 609–625.
- Daprati E., 1997. Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition* ***65***, pp. 71–86
- Eslinger P.J., Warner G.C., Grattan L.M. & Easton J.D, 1991. "Frontal lobe" utilization behavior associated with paramedian thalamic infarction. *Neurology*, **41** (3), pp. 450-452
- Frith C.D., Blakemore S.J. & Wolpert D.M., 2000. Abnormalities in the awareness and control of action, *Philosophical Transactions of the Royal Society of London B* **355**, pp. 1771–1788.
- Georgieff N. & Jeannerod M., 1998. Beyond consciousness of external reality. A 'Who' system for consciousness of action and self-consciousness, *Consciousness and Cognition* **7**, pp. 465–477.
- Gerrans P., 2002. Multiple paths to delusion, *Philosophy, Psychiatry, and Psychology* **9** (2002) (1), pp. 65–72.
- Haggard P., Clark S. & Kalogeras J., 2002. Voluntary action and conscious awareness, *Nat. Neurosci.* **5** (4), pp. 382-385.
- Jeannerod, M., Arbib, M.A., Rizzolatti, G. & Sakata, H., 1995. Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends Neurosci.* **18**, pp. 314–320.
- Jeannerod M., 1999. To act or not to act: Perspectives on the representation of actions, *Quarterly Journal of Experimental Psychology* **52A** (1999), pp. 1–29.
- Jeannerod M., 2003. The mechanism of self-recognition in humans, *Behavioural Brain Research* **142** (2003), pp. 1–15.
- LhermitteF., Pillon B. & Serdaru M., 1986. Human autonomy and the frontal lobes: I. Imitation and utilization behavior: A neuropsychological study of 75 patients. *Annals of Neurology* **19**, pp. 326–334.
- Mellors C.S., 1970. First-rank symptoms of schizophrenia, *British Journal of Psychiatry* **117**, pp. 15–23.
- Rizzolatti G., Camarda R., Fogassi L., Gentilucci M., Luppino G. & Matelli, M., 1988. Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. *Exp. Brain Res.* **71**, pp. 491–507.
- Rizzolatti G., Fadiga L., Gallese V. & Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* **3**, pp. 131–141.
- Sala S.D., Marchetti C. & Spinndler H. Right sided anarchic (alien) hand: A longitudinal study. *Neuropsychologia* **29** (11), pp. 1113-1127
- Sato A. & Yasuda A., 2005. Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*. 94(3), pp. 241-55.
- Spence S.A., 1996. Free will in the light of neuropsychiatry. *Philos. Psychiatry Psychol*. 3, pp. 75–90
- van den Bos E. & Jeannerod M., 2002. Sense of body and sense of action both contribute to self recognition, *Cognition* **85**, pp. 177–187.
- de Vignemont F. & Fourneret P., 2004. The sense of agency: a philosophical and empirical review of the "Who" system. *Conscious Cogn.* **13,** pp. 1-19.
- Wegner D.M. & Wheatley T., 1999. Apparent mental causation: sources of the experience of will. *Am. Psychol.* **54**, pp. 480–492
- Wolpert D.M., Ghahramani Z. & Jordan M.I., 1995. An internal model for sensorimotor integration. *Science* **269**, pp. 1880–1882.
- Wolpert D. M. & Kawato M., 1998. Multiple paired forward and inverse models for motor control. *Neural Networks* **11**, pp. 1317–1329.
- Wolpert D. M. & Ghahramani Z., 2000. Computational principles of movement neuroscience. *Nature Neurosci.* **3**(Suppl.), pp. 1212–1217.

PHYLOGENETICALLY ACQUIRED REPRESENTATIONS AND HYBRID EVOLUTIONARY ALGORITHMS

Adrianna Wozniak
Institute for Cognitive Sciences, Lyon, France
Phone 0033 4 37 91 12 17
wozniak@isc.cnrs.fr

First, we explain why Genetic Algorithms (GAs), inspired by the Modern Synthesis, do not accurately model biological evolution, being rather an artificial version of artificial, rather than natural selection. Being focused on optimisation, we propose two improvements of GAs, with the aim to successfully generate adapted, desired behaviour. The first one concerns phylogenetic grounding of meaning, a way to avoid the Symbol Grounding Problem. We give a definition of Phylogenetically Acquired Representations, based on a parallel between the notions of *representation* and of *adaptation*. In the second part of the paper, we propose a hybrid version of genetic algorithms, differently organizing the flow of genetic information by introducing inheritance of acquired traits and Horizontal Gene Transfer, a good tool for handle a cumulative directional process of artificial selection.

Genetic Algorithms as artificial versions of artificial, and not natural selection

Evolutionary Computation (EC) refers to methods for designing autonomous agents (artificial systems like physical or simulated robots, software agents) inspired by biological evolution, as the Modern Synthesis (MS) understands it. One of those methods is Genetic Algorithms (GAs). EC and GAs use biological ideas for two main purposes: optimisation and modelling.

Optimisation, because the evolutionary process by natural selection is identified with seeking for optimum, for good or best “solution” to the problem of reproduction and/or of survival of autonomous agents. It is an instrumentalist, pragmatist goal of AI: efficacy in creating agents

capable of successful operations relative to precise problems, in partially unknown environments without any intervention of the experimenter. AI uses artificial evolution because other methods are not successful (Harnad, 1990).

Modelling is the second and realist purpose underpinned by the hope that the better we know how reality works — given that reality works well — the more efficient our methods will be. The goal of AI, as those of other sciences, is to model and therefore to discover causal dependencies in evolutionary processes by natural selection. On the one hand, AI models and AI simulations are crudely simplified with regards to the heterogeneity of the evolutionary realm; on the other hand, GAs isolate the external causes and internal effects, thus having the advantage of leaving the possibility of grasping causal relations open to empirical investigation. Of course, even if models and simulations help to discover the existence of such causal relations, it doesn't imply either that the causal mechanisms discovered this way give rise to processes identical to those that occur in nature, nor that they are the only factors that take part in those processes.

GAs are considered as a formal study of adaptation, an artificial version of natural selection (Goldberg, 1989). According to MS, adaptation is a "*mechanism thanks to which external cause is transformed into effect*"(Lewontin, 2003:118-120), an asymmetrical process where « *the environment brings about an organic change exactly in its own image* » (Godfrey-Smith, 1996:86), and where "*organisms adapt to theirs environments, never vice versa*" (Williams, 1992:484). In MS, the movement of natural selection is environmentally driven (the environment differentiates between two genotypes G_1 and G_2). GAs follows this externalist concept of the phylogenetic relation environment/organism and uses the traditional concept of adaptation in which populations move relative to stable *selective environments* (Brandon, 1990:45) defined by experimenter. The survival of the fittest among all genotypes in the population is computed as follows: the experimenter tests through the fitness function the abilities of the agent to solve the problem (s)he is interested in. Then, genotypes are selected probabilistically according to their fitness scores, and enter the mating pool, which engenders the next generation. Individuals are copied according to their so-called *function values* (in EC) or *fitness function* (in MS). Function is an intuitive notion of "some measure of profit, utility or goodness that we want to maximize" (Goldberg, 1989:10).

GAs do show the power of natural selection, as MS understands it. Nevertheless, according to the Extended Theory of Evolution (ETE, John Odling-Smee et al, 2003; Day et al, 2003),

natural selection is not instantiated simply by an external factor: what constitutes selective factor is a resultant of both, environmental and organismic variables. In GAs the selective environment represents some externally fixed values, while in ETE, the organism defines the referential within which the selective environment is measured. The only constant valid in all system of reference is the viability criterion (note that viability does not imply the externalist view of adaptation, as defined by MS). In ETE, there are two variables in the frame of reference of the selective environment and the change in the value of one (organism) drives the change in the value of the other (environment), and inversely. Selection in this context designates simultaneous and reciprocal causality. This is the reason why Evolutionary Computing, inspired from the Modern Synthesis of the Theory of Evolution, is not an artificial version of natural selection, as claimed (Goldberg, 1989:10), but rather an artificial version of artificial selection. Artificial selection differs from natural selection in that in the former, the organism evolves according to some externally defined function, while in the later one organism modifies itself the fitness (and its function) and the selective factor that it is supposed to adapt to. The organisms do not phylogenetically track an external factor, contrary to MS where natural selection is an asymmetrical process of one way (passive) adaptation of organism to an environmental, independent value. In ETE organisms evolve without direct reference to some external factor; population tends not to the optimum (in correspondence to an externally defined task) but to the value that is a resultant of environmental and organismic properties.

Therefore methods used in Artificial Intelligence do not model well the evolution by natural selection. GAs make use of an externally defined fitness function, but natural evolutionary processes are not engineering operations of adaptation according to externally fixed demands. Yet, this may be why ETE models wouldn't be of use for efficient evolving computer systems. After all, experimenters do not want to obtain any results, but results for a more or less specific task. Current ideas of evolutionary robotics, inspired from biological evolution, are used precisely in the field of function optimisation, for engineering purposes. GAs are an example of artificial selection and do show the power of natural selection, where the latter instantiates external factors, which

experimenters judge important. Clearly, GAs are motivated by an optimisation purpose to improve the artificial selection of artificial, engineering-like evolution.

The goal of GAs is to successfully generate desired behaviour, adapted to an externally fixed demand. In the present paper we propose two improvements for GAs. The first part (§2) will concern a conceptual twist avoiding, in our opinion, the Symbol Grounding Problem by means of a phylogenetic grounding of meaning. The second part (§3) will concern two propositions of improvements of GAs through a different organisation of the flow of genetic information.

Phylogenetic grounding of meaning

One of the problem of AI is how the meaning of an external factor can be grounded, integrated, i.e. made intrinsic to the agent (Harnad, 1990). How can the experimenter make the agent understand the meaning of an external factor (symbol) s/he is interested in? Harnad's model of cognition is purely connectionist, top-down and symbolic, in the sprit of behaviourism, where names are connected to objects through invariant patterns in the sensory projections, learned through exposure and feedback. The meaning is supposed to be acquired via learning and is defined as a semantic correspondence with symbols. In this type of approach, the meaning of symbols emerges from the connection between the symbol system and the world (Fodor, 1994). Representational cognition is based on higher-order mental states and symbols (as Good Old Fashioned Artifcial Intelligence stated, Newell et al, 1976).

The AI definition of representation, as a direct mapping between internal symbols and external objects, has been undermined; nowadays learning is defined through interactions of the virtual individual with the world (Brooks, 1991). Therefore, behavioural responses join the rank of cognitive instances, though still only of those that are ontogenetically acquired[109]. The notions of learning and of adaptation are both localised at the ontogenetic level: learning mechanisms give the individual the ability of adapting to the environment and of elaborating behaviour in order to maintain itself in a viable state. Representations are learned (never hard-wired) and of belief-type; they acquire their function (meaning) through the ontogenesis where individuals learn what a given fact indicates; e.g. birds learn (in ontogenesis) that the Monarch butterfly

109 Ontogenical acquisition is acquisition that takes place during the individual's life.

marks indicate inedibility which leads them to an avoidance behaviour. Representations must be the causes of behaviour; in this sense, reflex processes over which the individual has no control are not representational or cognitive states. This is linked with the question of agency: I have cause to do this or that, but it is not for this reason that I am doing it (representations must be both reasons and causes of actions, Dretske, 1999). The reason is the belief and the belief is acquired through ontogenetic experience.

The current AI concept of representation— as learning during ontogenetic interaction with environment (Brooks, 1991)— misses one important fact, namely that ontogenetic learning is only one among two modes of meaning acquisition. The first one is obviously ontogenetic learning, where the individual acquires the meaning of x during its individual life. The second one is phylogenetic, where the individual benefits from the knowledge about the meaning of x acquired during the phylogenetic adaptation of the species it belongs to. For many researchers, cognitive states cannot be ascribed to phylogenetically acquired properties. For them, evolutionary adaptation or phylogenetic learning is different from "true" learning where changes in the behaviour are individually acquired during the ontogeny of the cell (Kilian and Muller, 2001).

Nevertheless, if learning means a modification of the internal states of an organism (or parameters in a virtual individual) during its interaction with the environment, learning does take place during individual experience *and* during species experience. What's more, learning mechanisms enabling ontogenetical adaptation of individuals to the environment and behaviour maintaining them in a viable state, already seat in their innate cognition, i.e. are based on phylogenetically acquired structures carried by *genetic open programs* (Mayr, 1974).

It is an old and plausible idea (developed by Platon[110]; Descartes[111]; Leibniz; Kant[112]; Lorenz[113]; Chomsky, 1975), that there is

110 Platon's (Socrates') methods of revealing by questioning (a slave boy), in the Meno.

111 "And man who rightly observes the limitations of the senses, and what precisely it is that can penetrate through this medium to our faculty of thinking must needs admit that no ideas of things, in the shape in which we envisage them by thought, are presented to us by senses. So much so that in our ideas there is nothing which was not innate in the mind, or faculty of thinking". Quoted in Chomsky, 1975.

112 "(…) what is borrowed solely from experience is, as we say, known only a posteriori, or empirically. Now we find, what is especially noteworthy, that even into our experience there enter modes of knowledge which must have their origin a priori, and which perhaps serve only to give coherence to our sense-representations. For if we eliminate from our experience everything which belongs to the senses, there still remain certain original concepts and certain judgments derived from them, which must have arisen completely

nothing in the representation, which does not come from the sensory, individual experience, except the senses, the cognitive apparatus itself[114]. The evidence and the measure for phylogenetically acquired and (partially) innate components of cognitive and representational states would be the following: if we take sensory experience as the input and behavioural response of the individual as the output, we will see that the output contains more information than provided by the individual, sensory experience of external stimulus. We subtract the stimulus from the output; we thus obtain the contribution brought by innate knowledge. It brings out the fact that representation contains an innate component, and pinpoints the existence of an innate cognitive endowment of the organism. If representations are underpinned by innate components in such a way that the latter are indispensable for those representations, the innate components also must be considered as part of the representation.

Obviously, many innate cognitive and representational states are not fully manifested at birth, and the presence of some external, triggering, factor is needed for these ideas to become available (Ariew, 1996; Lorenz, 1966). Thus, the representation of the world is built not only from learned components, but depends also on the innate ones. Evolving organisms benefit from the combination of phylogenetic and ontogenetic learning. It raises a few points against the exclusivity of intentional conceptions: why do we attribute representational status to ontogenetically acquired features but refuse it to hard-wired ones? There is a striking parallel between the notions of *representation* and of *adaptation*, that will lead us to the notion of Phylogenetically Acquired Representations (PAR):

Representation	**Adaptation**
An (a set of) internal state(s) of the agent	A (a set of) hereditary (partly carried by open genetic program) property of the

a priori, independently of experience, inasmuch as they enable us to say, or at least lead us to believe that we can say, in regard to the objects which appear to the senses, more than mere experience would teach – giving to assertions true universality and strict necessity, such as mere empirical knowledge cannot supply”. (Kant, 1781:A2)

113 “(…) the blueprint contained in the genome requires innumerable environmental factors in order to be realised in the phenogeny of structures and functions. During his individual growth, the male stickleback may need water of sufficient oxygen content, copepods for food, light, detailed pictures on his retina and millions of other conditions in order to enable him, as an adult, to respond selectively to the red belly of rival. Whatever wonders phenogeny can perform, however, it cannot extract from these factors information which simply is not contained in them, namely, the information that a rival is red underneath”. (Lorenz 1966:37)

114 Paraphrase de Leibniz: Nihil est in intellectu, quod non fuerit in sensu, excipe: nisi ipse intellectus.

	agent
that holds a relation of reference	that results from a causal phylogenetic relation
toward certain objects in the external world.	toward an external factor.
The representation of the object, as present in the mind, does not entirely derive from sensory, individual experience of this object.	The contribution of the sensory, individual experience of this factor is not sufficient for the trait-adaptation to develop.

PARs as adaptations. PARs are (a set of) features of the organism carried by open genetic programs that result from a causal phylogenetic relation with factors from the selective environment. The forms of PARs are thus not entirely determined by individual experience of the environmental factor.

PARs as representations. Phylogenetically acquired features have representational status, because adaptations (e.g. adaptative escape behaviour) corresponding to an environmental factor (e.g. snake), do not derive and cannot be fully explained, by the ontogenetically acquired experience of this factor. The ontogenetical exposure to snakes is not sufficient to acquire the escape behaviour that is triggered once the individual senses a snake. The reason for which individuals of species S fly snakes is not an ontogenetically acquired belief of these individuals, but precisely a PAR, the meaning of a snake being acquired through the phylogenetic experience of S.

Natural selection is a process of discriminating sampling occurring when the individuals do not reproduce because their traits does not fit to their environment. The chance of individuals to contribute to the next generation depends on this fitness. In GAs natural selection designates a cause/effect relation, whereby the environment (as a fitness value fixed by experimenter) instantiates the cause and the organism instantiates the effect. This causal and externalist characteristic of natural selection guarantees that the main criterion of representation is fulfilled, namely the presence of the causal relation from object to representation. Thus, PAR is every feature that constitutes an adaptation, i.e. resulting from the discriminating process of natural selection. Since the latter can act only on what is heritable, and what is heritable is genetic, a structure that constitutes an adaptation must be (partially) innate[115]. There are three conditions for a feature F to be considered as representing x:

- F must enter the state S if x occurs, e.g. trigger escape behaviour in the presence of a sensory experience invoking a predator;

115 Not every innate trait has to be an adaptation.

F must be an adaptation:

- the property of F to enter the state S if x must be the cause thanks to which F was retained in the discriminating process of natural selection
- F must be underpinned by the open genetic program (innate to some extent)

How then can the concept of phylogenetic acquisition of meaning and the definition of Phylogenetically Acquired Representations help to solve the Symbol Grounding Problem? How can the meaning of an external factor be grounded, integrated in the agent? Meaning is supposed to be acquired via the phylogenetic process of natural selection (species learning and not only ontogenetic learning) and designates an adaptive (and not ontogenetically semantic) correspondence with external factors. The meaning of those factors emerges from the selective relation between them and the genetic program of the species. Representational cognition is based not on higher-order mental states and symbols but on partly innate features underpinning them. How then can the experimenter make the agent understand the meaning of an external factor s/he is interested in? We propose to take into account phylogenetic grounding, based on the assumption that the features-adaptations are rightfully representational ones and bear the meaning of the external factor according to which they evolved.

Hybrid Genetic Algorithms

In this part of the paper we will propose some ideas as to how to organise the flow of genetic information, rendering it more efficient in order to successfully generate the desired, adapted behaviour. To generate an evolutionary process, the three following requirements must be fulfilled. The first one is the *principle of variation*, i.e. the existence of polymorphism in morphologic, physiologic or behavioural traits within populations. At least some variants must be hereditary – *principle of heredity* – i.e. in the progeny's generation there must exist traits similar to those present in the parental generation. Without heredity, adaptive evolution is not possible (Dawkins, 1982), for only traits possessing genetic basis can be selected and passed from one generation to the next, and become an adaptation. Genes guaranty the possibility of transmission of selected variants. Finally, the *principle of selection*, is driven by fitness differences in the situation where some individuals, bearers of modified traits, leave more descendants than others. That is all we need to generate an evolutionary process of artificial selection. GAs

not only fulfil all those three necessary conditions, but also take, what is more, some additional ones that have come with relatively recent discoveries integrated in the Modern Synthesis. In the case of the principles of variation, MS states that variation has two sources, mutation and recombination. When it comes to the principle of heredity, GAs' models are constructed according to the *Central Dogma* of molecular biology setting out that DNA causes the production of RNA that makes proteins and then cells. The reverse process doesn't occur: proteins or cells don't determine on their turn the nucleic acid. The fact that genotype affects phenotype and that phenotype does not affect genotype implies that acquired traits do not affect an organism's genome and that only genome (and not what parents learned or acquired during their ontogenesis) is passed to the offspring. Genetic material is transferred to another organism that is a descendant, i.e. from parent to offspring, in an intragenerational way. This is called *vertical gene transfer* (VGT).

However, all those conditions are additional to the three ones necessary to generate an evolutionary process of artificial selection. Why are they accessory? Darwin developped his theory of natural selection (in 1859) without knowing exactly either the source of variation or the nature of inheritance. Before him, in 1809, Lamarck proposed his concept of evolution, where variation is somehow induced by the environment (variation is neither spontaneous nor random, as in MS), and the parental organisms transmit to their offspring the traits that they acquired in ontogenesis (contrary to the Central Dogma of MS). The mechanisms generating variation and responsible for inheritance were known much later. In 1866, Mendel gave the basis for the understanding of genetic recombination, and in 1904 Weismann showed that the germ line is segregated from the soma, thanks to the observation that the offsprings of mice with cut-off tails have normal tails. The conviction about the genome as a *one way transducing device* was reinforced after 1958 with the discoveries in molecular biology of Watson and Crick.

The goal of GAs is to successfully generate desired behaviour, adapted to an externally fixed demand. More realistic and complex genetic algorithms were conceived in order to obtain a precise result. Many evolutionarily inspired tricks were incorporated at different levels, like genetic transfer during cellular division (inversion, translocation, deletion, etc.), diploidy and sexual reproduction, coevolution (host-parasite, arm races), sexual selection, etc. MS inspired all those models. Nevertheless, VGT is a kind of frozen accident, far from being universal (its exceptions are e.g. retroviruses, retrotransposons, prions). "The non-inheritance of acquired characters is a contingent fact, usually but not always true, not a logical necessity" (Maynard-Smith, 2001). The same is

valid for the source of variation. To generate an evolutionary, selective process, there must exist heritable variants and factor differentiating them, but the way of generating and making those variants inherited does not need to be exactly copied from nature. It can be even simpler and maybe more efficient for engineering and optimisation purposes. We will now propose bipartite candidate theoretical solution, which we call Hybrid Genetic Algorithms (HGAs), for the current state of technology can provide experimental tools following this conceptual liberty.

Acquisition of acquired traits

In current models of GAs, acquired traits do not affect an organism's genome, which has some important implications. First, at least one generation is needed for the adaptative process to take place. Desirable combination (coming from intra-chromosomal or inter-chromosomal recombination) or an advantageous mutation can be simply lost and do not appear in the next generation. It is a drawback of the intragenerational mode of transmission that the (advantageous) variant traits must be generated de novo in each generation. The further implication of VGT is that what individuals learn during their lifetime is not genetically transmitted to the next generation. This is due to the fact that the ontogenetically acquired characteristics are not directly copied to the next generation, but the genes underpinning them. Consequently, the ontogenetic increase in performance relative to the fitness function is lost at the end of the individual life. AI can create evolutionary processes that function in a simpler manner and where the selective retention of adaptative traits, including those acquired during ontogenetic learning, is possible. In HGAs, it is not only genotype that would affects phenotype, but phenotype could also affect genotype. For example, in a robot controlled by an artificial neural network, genome would modify synaptic weights, as before, and additionally this change would directly drive a change in the genome. The adaptation would trigger an ontogenetic (and not phylogenetic) modification of the genome, a horizontal heritable trait acquisition. HGAs would take a Lamarckian orientation and acquired (learned) traits of an individual would affect its genome. The ontogenetic increase in performance according to the fitness function wouldn't be lost.

Thanks to inheritance of acquired features, an advantageous propriety that an individual acquires in the process of learning will be transmitted to the next generation. For instance, an individual in a population P learns something about the object x, vitally related to all individuals of P. This knowledge allows this individual to progress

(according to the fitness threshold established by the experimenter) and to gain further knowledge of x.

Acquired DNA or Horizontal Gene Transfer (HGT)

Once we have at our disposition horizontal heritable trait acquisition we can enrich the method with horizontal gene acquisition. Suppose that the experimenter would like to spread among all individuals of the population ontogenetically gained feature and then encod it in the genome. In order to do it with GAs' methods, s/he must apply directional selection and wait a number of generations to see the desired effect universally fixed. However, there is a possibility to make the desired trait horizontally displaceable by introducing to the model the exchange of the genetic material within generation (interspecific recombination without creating new individuals). This genetic free swapping within population could be made by introducing Horizontal Gene Transfer (HGT), characteristic of the evolution of the cell before early, primitive cells differed in three primary lines of descent: bacteria, archea and eukaryotes (before Darwinian threshold, Ochman and all, 2000: 304). In HGT, substantial amounts of DNA are introduced (or deleted) from the chromosome. HGAs models would resemble a kind of mosaic or net, metaphors visualising the HGT exchange occurring at the roots of the tree of life. This would be a tool for the experimenter to improve the process of cumulative and directional selection.

In HGAs, population would be considered as a universal genetic pool, and HGT as a way of redistributing desired (non desired) traits. This would multiply the range of combinatorial heritable possibilities and increase the chance of obtaining the trait the experimenter is interested in. The content and the structure of genomes in the population, moulded by HGT, would probably display a wide degree of variants what would enable phylogenetic plasticity and increase the chance to obtain the desired characteristic[116].

New traits would appear not only after point mutations or genetic recombination (intra-chromosomal — combination of parental and maternal genes— or inter-chromosomal — of chromosomes), but also to interspecific recombination[117], possible thanks to HGT. All desirable

---

116 HTG explains why bacteria develop their incredible antibiotic resistance, their ability to adapt to the environments.

117 In the nature, DNA sequences are even transferred among taxa, being acquired from distantly related or non related organisms, e.g. Adzuki Bean Beetle's genome contains some sequences from the genome of Wolbachia, its parasit.

novelties (acquired during the ontogenetic learning, due to the point mutations, etc.) could be shared and henceforth evolve simultaneously. This would create an *unlimited system of heredity* (Maynard Smith and Szathamary, 1995), where a trait can vary into a great number of heritable states, as in the case of prokaryotes and bacteria or of languages and cultures. Of course, as in the vertical mode of acquisition, natural selection (i.e., the experimenter) is the arbiter of the adaptive value of traits.

Conclusions

In the first part of the paper, we explain the conceptual revolution made by the Extended Theory of Evolution (ETE, John Odling-Smee et al, 2003; Day et al, 2003). The latter points out that selective environment and fitness value, according to which organisms are supposed to evolve, are a resultant of two variables, environment *and* organism. Natural selection is not a simple externalist relation; the organisms do not only evolve in response to an external factor, but themselves partly define the fitness function. Thus, GAs are an instantiation of artificial selection, whose main purpose is optimisation, not realistic modelling.

We thus propose two improvements, conceptual and technical, in generating a desired, adapted behaviour. The first one (§2) concerns the phylogenetic grounding of meaning, a way to avoid the Symbol Grounding Problem. We explain the parallel between the notions of representation and of adaptation and elaborate the concept of Phylogenetically Acquired Representation. The meaning of an external factor can be grounded, integrated to the agent via the process of artificial selection; we take seriously the phylogenetic mode of acquisition, species learning, and consider feature-adaptations as legitimate representational ones, bearing the meaning of the external factor according to which they evolved.

In §3, we propose Hybrid Genetic Algorithms (HGAs), a melange of real and fictitious elements of evolutionary processes. We propose to incorporate to GAs horizontal heritable trait acquisition (inheritance of acquired traits) enriched by horizontal gene acquisition, a tool for the experimenter to handle the cumulative directional process of artificial selection. It introduces, in comparison to the Modern Synthesis, additional evolutionary mechanisms:

- new source of variation (Horizontal Gene Transfer, HGT, makes possible intraspecific recombination) and

- additional modes of inheritance, enabling the experimenter to easily conserve and spread or delete selected features (HGT) and ontogenetical modification of the genome (inheritance of acquired traits), contributing to the gain of the performance according to the fitness function.

Acknowledgments

I thank Dr. Anne Reboul for her feedback and support, and Regis Martinez for his comments on Hybrid Genetic Algorithms.

References

Ariew, A. 1996. Innateness and canalization. In *Proceedings in Philosophy of Science,* Cleveland, Ohio. pp. 19-27.

Brandon, Robert N. 1990. *Adaptation and environment*. Princeton University Press.

Brooks, R. 1991. Intelligence without reason. *Proceedings of 12th Int. Joint. Conf. On Artifcial Intelligence*, Sydney, Australia. 569-595.

Chomsky, Noam. 1975. Cartesian linguistics: acquisition and use of language. In *Innate ideas*. St. Stich (ed). University of California Press. Berkeley, Los Angeles, London.

Dawkins, Richard. 1982. *The extended phenotype. The long rich of the gene*. Oxford University Press, New York.

Day, Rachel; Laland, Kevin et Odling-Smee, John. 2003. Rethinking adaptation: the niche-construction perspective. *Perspectives in Biology and Medicine*, vol. 46, n° 1. pp. 80-95.

Dretske, F. 1999. Machines, Plants and Animals: The Origins of Agency. *Erkenntnis* 51: 19–31.

Fodor, Jerry. 1994. *The Elm and the Expert, Mentalese and its Semantics*. The 1993 Jean Nicod Lectures. MIT Press.

Godfrey-Smith, Peter. 1996. *Complexity and the function of the mind in nature*. Cambridge University Press.

Goldberg, D. E. 1989. *Genetic Algorithms is Search, Optimisation and Machine Learning*. Boston: Addison Wesley.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42 :335-346.

Kant, Immanuel. 1781. *Critique of pure reason*. First edition.

Kilian A. E. and Muller, B. S. 2001. Life-like Learning in Technical Artefacts: Biochemical vs. Neuronal Mechanisms. Web address: http://www.ais.fraunhofer.de/%7Ebsm/CR1714.pdf.

Leibniz, Gottfried Wilhelm von. New Essays on Human Understanding. http://www.earlymoderntexts.com/pdfbits/leibne.html

Lewontin, Richard C. 2003. *La triple hélice. Les gènes, l'organisme, l'environnement*. Editions du Seuil. Titre original : *Gene, organismo e ambiente*. Editeur original : 1998, Gius, Laterza & Figli Spa, Roma-Bari.Lorenz, Konrad. 1966. *Evolution and Modification of Behaviour*. Methuen and Co., London.

Maynard Smith, J and Szathamary, E. 1995. *The major Transitions in Evolution*. Oxford: Oxford University Press.

Mayr, Ernst. 1974. Behavior Programs and Evolutionary Strategies. *American Scientist* 62: 650-659.

Newell, A., Simon, H. A. 1976. Computer science as empirical inquiry: Symbols and search. *Commun. Assoc. Comput. Machinery* 19 () 111-126.

Ochman Howard; Lawrence, Gefrey and Groisman, Eduardo. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, vol. 405. p. 299-304.

Odling-Smee, John ; Laland, Kevin et Feldman, Marcus. 2003. *Niche construction : the neglected process in evolution*. Princeton University Press

Willliams, G.C. 1992. Gaia, nature worship, and biocentric fallacies. *Quarterly Review of Biology*. 67 :479-486.

WHY COMPUTERS CAN'T FEEL PAIN

Mark Bishop

Department of Computing

Goldsmiths College, University of London, UK

m.bishop@gold.ac.uk

Background

In Science and Science Fiction the hope is periodically reignited that a computer system will one day be conscious in virtue of its execution of an appropriate program; indeed the UK funding body EPSRC recently awarded an 'Adventure Fund' grant of around £500,000 to a team of 'Roboteers and Psychologists' at Essex and Bristol universities[118], with a goal of instantiating machine consciousness - in a 'humanoid-like' robot called Cronos - through appropriate computational 'internal modelling'. What I will outline below is a brief reductio style argument that either suggests such optimism is misplaced or that panpsychism - the belief that the physical universe is composed of elements each of which is conscious - is true.

First though, it is helpful to outline exactly what is meant by the term consciousness in the context of this paper. By phenomenal consciousness I refer to the first person, subjective phenomenal states -

118 The project, 'Machine consciousness through internal modelling', is funded by the EPSRC Adventure Fund. The total funding is £493,000 split between the departments of Computer Science, University of Essex and the Department of Psychology, University of Bristol. The project is led by Professor Owen Holland, University of Essex.

sensory tickles, pains, visual experiences and so on. Current research into perception and neuro-physiology suggests that physically identical brains will instantiate identical phenomenal states; however, pace Maudlin (1989), “if a causal theory of reference is correct, a molecule-for-molecule identical replica of my brain, if just brought into existence, may not be capable of entertaining the proposition that ice is made of water. Still our best guess is that such a brain would support identical states of consciousness to mine, identical phenomenal states.” As Maudlin (ibid) observes, this thesis is not analytic; however something like it underpins computational theories of mind; for computational structure supervenes on physical structure – physically identical brains must be computationally identical. Hence Maudlin (ibid) formulates the ‘supervenience thesis’, “two physical systems engaged in precisely the same physical activity through a time will support precisely the same modes of consciousness (if any) through that time.”

The core argument I outline in this paper derives from ideas originally outlined by Hilary Putnam (1988), Tim Maudlin (1989) and John Searle (1990) and subsequently criticised by David Chalmers (1996), Colin Klein (2005) and Ron Chrisley (2006) amongst others[119]. In what follows, instead of seeking to justify Putnam’s claim that “every open system implements every Finite State Automaton (FSA)”, and hence that psychological states of the brain cannot be functional states of a computer, I will seek to establish the weaker result that, over a finite time window, every open physical system implements the trace of a Discrete State Machine Q, as it executes its control program on fixed, specified input (x). That this result leads to panpsychism is clear as, equating Q (x) to a specific computational system that is claimed to instantiate phenomenal states as it executes, and following Putnam’s procedure, identical computational (and ex-hypothesi phenomenal) states can be found in every open physical system.

The route-map for this endeavour is as follows. In the first part of the paper I introduce Discrete State Machines, DSMs, and show how, with input to them defined, their behaviour is described by a simple unbranching sequence of state transitions analogous to that of an inputless DSM. Then I review Putnam’s 1988 argument that purports to show how every open physical system implements every inputless FSA. This argument is subsequently applied to a robotic system that is claimed to instantiate genuine phenomenal states as it operates. The paper

119 Cf. Minds and Machines, 4: 4, ‘What is Computation?’, November 1994.

concludes with a brief discussion of some objections raised following presentation of these ideas at the "Computers and Philosophy" conference, Leval 2006.

Discrete State Machines

In his 1950 paper, 'Computing Machinery and Intelligence', Turing defined Discrete State Machines, DSMs, as "machines that move in sudden jumps or clicks from one quite definite state to another", and explained that modern digital computers fall within the class of them. An example DSM from Turing is one that cycles through three computational states (Q_1, Q_2 & Q_3) at discrete clock clicks. Turing demonstrated that such a device, which cycles through a linear series of state transitions 'like clockwork', may be implemented by a simple wheel-machine that revolves through 120^0 intervals.

By labelling the three discrete positions of the wheel $\{W_A, W_B, W_C\}$ we can map computational states of the DSM (Q_1, Q_2, Q_3) to the physical positions of the wheel $\{W_A, W_B, W_C\}$, such that, for example, ($W_A \Rightarrow Q_1$; $W_B \Rightarrow Q_2$; $W_C \Rightarrow Q_3$). Clearly this mapping is observer relative: position W_A of the wheel could equally map to computational states Q_2 or Q_3 and, with other states appropriately assigned, the machine's state transition sequence (and hence its function) would remain unchanged. It is central to the argument to be developed in this paper that **all** computational states are observer relative in this fashion; they are not intrinsic to the physics of the system – that is, their determination always involves an 'observer-specified' function that maps from physical system state onto computational state.

In general, we can generate the behaviour of any K-state (inputless) DSM, (f (Q) $\Rightarrow$ Q'), by a K-state wheel-machine (e.g. a digital counter), and a function that maps each wheel/counter state W_n/C_n to each computational state Q_n as required.

In addition, Turing's machine may be stopped by the application of a brake and whenever it enters a specific computational state a lamp will come on. Input to the machine is thus the state of the brake, (I = {ON | OFF}), and its output, (Z), the state of the lamp. Hence the operation of a DSM with input is described by a series of 'contingent branching state

transitions', which map from current state to next state, f (Q, I) => Q' and define the machines output - in the Moore form - as f (Q') $\Rightarrow$ Z.

However, (over a finite time interval), defining the input to the DSM entails that such 'contingent behaviour' reverts to 'clockwork', (f (Q) => Q'). E.g. If Turing's DSM starts in Q_1 and the brake is OFF for two clicks, its behaviour, (execution trace), is fully described by the sequence of state transitions, (Q_1, Q_2, Q_3); conversely if Turing's DSM starts in Q_1 and the brake is ON for two clicks, its behaviour - execution trace - is described by the sequence of state transitions, (Q_1, Q_1, Q_1).

Hence, over a finite time window, if the input to a DSM is defined, we can map from each wheel/counter state W_n/C_n to each computational state Q_n, as required. In Bishop (2002) I demonstrated, pace Putnam, how to map any computational state sequence with fixed [defined] input onto the [non-repeating] natural state sequence generated by any open physical system.

Putnam's mapping

Discussed in a brief appendix to Hilary Putnam's 1988 book *Representation and Reality* is a short argument that endeavours to prove that every open physical system is a realisation of every abstract Finite State Automaton and hence that functionalism fails to provide an adequate foundation for the study of the mind.

Central to Putnam's argument is the observation that every open physical system, S, is in different 'maximal' states[120] at every discrete instant and hence can be characterised by a discrete series of non-cyclic natural state transitions, [s_1, s_2 .. s_t .. s_n]. Putnam argues for this on the basis that every such open system, S, is continually exposed to electromagnetic and gravitational signals from, say, a natural clock. Hence by quantizing these natural states appropriately, every open physical system can by considered as a generator of discrete non-repeating modal state sequences, [s_1, s_2 .. s_∞][121].

120 A 'maximal' state is a total state of the system, specifying the system's physical makeup in absolute detail.

121 Chalmers (1996) observes, "Even if it [the claim that 'every open physical system is a realisation of every abstract Finite State Automaton'] does not hold across the board (arguably, signals from a number of sources might cancel each other's effects, leading to

Considering Turing's inputless DSM state machine, Q, and a six state digital counter [c_1 ... c_6], it is trivial to observe that, over time interval [t_1 .. t_6], if we map the state [Q_1] to the disjunction of counter states, [c_1 v c_4], DSM state [Q_2] to the disjunction of counting machine states, [c_2 v c_5] and DSM state [Q_3] to the disjunction of counting machine states, [c_3 v c_6], then the counting machine will fully implement Q as it transits counter states [c_1 .. c_6] over time interval [t_1 .. t_6]. Further, given any [counting] machine state, say [Q_1] $\in$ {c_1, c_4}, at time [t_1], we can modally predict that the DSM will enter state [Q_2] at time [t_2].

To show that being in state [Q_1] at time [t_1] caused the counter to enter state [Q_2] at [t_2] we observe that at [t_1] the counter is in state [c_1], (which the mapping function labels DSM state [Q_1]), and that being in state [c_1] at [t_1] causes the counter to enter state [c_2], (which the mapping function labels DSM state [Q_2]) at [t_2]. Hence, given the current state of the counter at time [t], we can predict its future state and hence how the states of DSM Q evolve over the time interval under observation.

Note, after Chalmers, that the counter-machine described above will only implement a particular execution trace of the DSM[122] and Chalmers remains unfazed at this result because he states that inputless machines

---

a cycle in behaviour), the more limited result that every non-cyclic system implements every finite-state automaton would still be a strong one".

122 Clearly there may be other state transition sequences that have not emerged in this execution trace. To circumvent this problem and fully implement an inputless FSA by an infinite state [counter] system, Chalmers posits the system with an extra dial - a sub-system with an arbitrary number of states, [c[dial-state, counter-state]]. Now, associate dial-state [1] with the first run of the FSA. The initial state of the counter machine will thus be [c[1, 1]] and we associate this with an initial state of the FSA. Next associate counter states [c[1, 2]], [c[1, 3]] with associated FSA states using the Putnam mapping described earlier. If at the end of this process some FSA states have not come up, we choose a new FSA state, [C], increment the dial of the counting machine to position [2] and associate this new state [c[2, 1]] with [C] and proceed as before. By repeating this process all of the states of the FSA will eventually be exhausted. Then, for each state of the inputless FSA there will be a non-empty set of associated counting machine states. To obtain the FSA implementation mapping we use Putnam's mapping once more and the disjunction of these states is mapped to the FSA state as before. Chalmers remarks, "It is easy to see that this system satisfies all the strong conditionals in the strengthened definition of implementation [above]. For every state of the FSA, if the system is (or were to be) in a state that maps onto that formal state, the system will (or would) transit into a state that maps onto the appropriate succeeding formal state. So the result is demonstrated." (Chalmers 1996, p.317). However this extension is not required for the argument developed herein.

are simply an “inappropriate formalism” for a computationalist theory of mind[123].

Clearly the addition of input makes the DSM formalism non-trivial. There can now be branching in its execution trace, as the next state is contingent on both its current state and the input. This gives the system a combinatorial structure. But, as Chalmers observes, Putnam’s revised construction does not properly encapsulate this structure – rather it merely manifests one trace of the FSA with a specific input/output dependency. So we are left with the counter intuitive notion that, for example, when using a rock to implement a two plus two program, we mark two on the input area of the rock and four on the output and credit the rock with computing the result..

In his 1996 paper, Chalmers introduces a more suitable FSA formalism, which makes explicit such input/internal-state dependencies, the Combinatorial State Automaton, CSA. A CSA is like - and no more powerful than - a conventional Finite State Automaton, FSA, except that its internal states, [S], are structured to form a set, $\{s_1, s_2 \ldots s_n\}$, where each element $\{s_i\}$ can take on one of a finite set of values or sub-states and has an associated state transition rule.

Chalmers then demonstrates how to map a CSA onto a physical system in such a way as to deal with such input/internal-state dependencies correctly and preserve the internal functional organisation of the original program, but only at the price of a combinatorial increase in the number of states required for the implementation. In fact, as he illustrates in his paper, executing even the most trivial FSA with input and output, over a small number of time steps would rapidly require a physical system with more states than atoms in the known universe to implement it. So it seems that “we can rest reasonably content with the knowledge that the account as it stands provides satisfactory results within the class of physically possible system”, and functionalism is preserved.

---

123 “To see the triviality, note that the state-space of an inputless FSA will consist of a single unbranching sequence of states ending in a cycle, or at best in a finite number of such sequences. The latter possibility arises if there is no state from which every state is reachable. It is possible that the various sequences will join at some point, but this is as far as the ‘structure’ of the state-space goes. This is a completely uninteresting kind of structure, as indeed is witnessed by the fact that it is satisfied by a simple combination of a dial and a clock. (ibid., p.318).

The problem that the CSA makes explicit is that of fully encapsulating the complex inter-dependencies between machine state and the input. To implement these using an open physical system requires an astronomical number of internal states, whereas the simple implementation of an inputless FSA that Putnam describes functions only because of the subsequent loss of generality. However, as we observed with Turing's DSM, when input is defined over a specific time interval, the combinatorial state structure collapses to a bounded linear path which can be simply generated using Putnam's mapping and any open physical system.

Returning to a putative conscious robot such as Cronos; at the heart of such a beast there is a computational system – typically a microprocessor; memory and memory mapped peripherals. Such a system forms a Discrete State Machine, DSM in interaction with its environment[124]. Thus, recalling that the computational states of DSMs are 'observer-relative' - requiring a mapping function to be fully determined from the physical state of the system - we note that with input to the robot specified and fixed over a finite time interval, we can simply map the execution trace of its control program onto the state evolution of any digital counter (or, pace Putnam, any open physical system).

Hence, if the state evolution of the robot DSM instantiates phenomenal experience, then so must the state evolution of any open physical system and we are inexorably led to embrace a panpsychist worldview where phenomenal consciousness is found everywhere.

Objections: (1) Do counterfactuals matter?

In Bishop (2002) I discuss several objections to this reductio with, perhaps, the most potent coming from David Chalmers who argues that 'as the above only implements one execution trace of the DSM it is not sensitive to counterfactuals; and it is only the possibility of appropriate counterfactual behaviour that guarantees phenomenal experience'

124 NB. It is central to the computationalist underpinning of cronos that its putative conscious states are not contingent upon it physically interacting with a physical environment; in personal communication, Prof. Holland envisaged a possible follow up project in which the entire cognitive architecture of cronos and its environment are entirely implemented in software, in a large scale virtual reality simulation.

My initial response to this line of argument (Bishop 2002a; Bishop 2002b); followed from Maudlin's Supervenience thesis. Consider what happens if a putatively conscious robot, R_1, with full counterfactual sensitivity, is step-by-step transformed into new robot R_2, such that its resulting behaviour is determined solely by a linear series of state transitions; substituting each conditional branching state transition sequence in the evolution of R_1, with a linear state transition defined by current state and the defined input. It seems clear that, over a finite time interval and with identical input, the phenomenal experience of R_1 and R_2 must be the same. Otherwise we have a robot, R_n, ($R_1 < R_n \leq R_2$), whose phenomenal experience is somehow contingent upon the presence or absence of non-entered state sequences contravening Maudlin's 'supervenience thesis' (outlined earlier)[125]. However at the 2006 Tucson consciousness conference, in a paper entitled 'Counterfactual computational vehicles of consciousness', Ron Chrisley suggested that as we morph between R_1 and R_2, with the deletion of each conditional non-entered state sequence real physical differences between the robots emerge. Effectively, with each replacement of each of the non-entered conditional state sequences, we crucially no longer execute their concomitant conditional test and branch instructions[126]; hence the core reductio no longer holds.

To address this criticism I will endeavour to illustrate that the mere execution of a conditional branch instruction where the result of the test is known and fixed also cannot affect any putative phenomenal states instantiated by the program.

Some conditional branch instructions:

- IF (A > B> THEN GOTO {statement sequence A} ELSE {B}
- IF (A > 10) THEN GOTO {statement sequence A} ELSE {B}
- IF (11 > 10) THEN GOTO {statement sequence A} ELSE {B}

---

125 "Suppose that a system exists whose activity through a period of time supports a mode of consciousness, e.g. a tickle or a visual sensum. The supervenience thesis tells us that, if we introduce into the vicinity of the system an entirely inert object that has absolutely no causal or physical interaction with the system, then the same activity will support the same mode of consciousness. Or again, if the activity of a system supports no consciousness, the introduction of such an inert and causally unconnected object will not bring any phenomenal state about … if an active physical system supports a phenomenal state, how could the presence or absence of a causally disconnected object effect that state?" (Maudlin, 1989).

126 A 'conditional branch' instruction is an instruction in a computer program of the form, "IF (TEST IS TRUE) THEN GOTO {statement sequence A} ELSE GOTO {statement sequence B}".

A non-conditional branch instruction:

- GOTO {statement sequence A}

The first conditional branch simply states that IF the value of variable A is greater than that of variable B then execute statement sequence {A} otherwise execute statement sequence {B}. The second conditional is of the same form, however this time we are comparing the value of variable A with the literal value '10'. However in the third example, the 'conditional' compares the value of two literals (11 and 10), hence the result of the test will always be true and the program will always follow statement sequence {A}. The fourth example is of a simple branch instruction, whereby control of the program will unconditionally shift to statement sequence {A}.

At this juncture it is critical to note that many modern 'optimising compilers' will automatically convert the third conditional statement to a simple branch instruction (as these execute more efficiently). Further, if the compiler can deduce that the value of A can never be less than or equal to ten during any possible execution of the program, an optimising compiler may also convert the second conditional into a simple branch; similarly, if it can be deduced a priori that A is always going to be greater than B then it may even convert the first statement into a simple branch; hence it is clear that no special phenomenal properties can result from the mere execution of a conditional statement, otherwise the phenomenal properties of a putative robotic system would be in a strong sense conditional on the type of compiler used to compile its control program.

I will now describe four segments of code, used in four, otherwise identical, robots [A .. D], each of which has a red Munsell colour card placed in view of its optical sensor. Electronic circuitry ensures that the value registered by the optical sensor is stored in a digital latch circuit, positioned at location $FFFF[127] in the computer's memory. If, say, the colour sensor indicates red light falling on it, it will register say $FF, otherwise, if say it is in darkness, it will register say $00.

ROBOT A: Sensor reading genuinely contingent on the current ambient light conditions.
LDA $FFFF

127 The $ sign indicates a hexadecimal number; i.e. a number to the base 16; dgit range is [0 .. 9 A .. F], hence hexadecimal $FF is 15 x 16 + 15 = 255 (decimal).

IF (A = 0) THEN execute statement sequence {A} ELSE {B}

ROBOT B: Red light permanently illuminates the sensor, so it always registers $FF and $FF is always stored by the latch at location $FFFF.
LDA $FFFF
IF (A = 0) THEN execute statement sequence {A} ELSE {B}

ROBOT C: Sensor faulty so it always registers $FF hence $FF is always stored by the latch at location $FFFF.
LDA $FFFF
IF (A = 0) THEN execute statement sequence {A} ELSE {B}

ROBOT D: The latch is forced to always store $FF at location $FFFF; hence the value subsequently loaded from $FFFF will always be $FF.
LDA $FFFF
IF (A = 0) THEN execute statement sequence {A} ELSE {B}

The question for the computationalist roboteer is which of the four robots [A .. D] will experience phenomenal red. It would appear that, ex-hypothesi, robot A must experience red, as the value obtained from the latch is an accurate reflection of the light signal falling on the sensor. By similar logic, robot B must also experience phenomenal red. – if a different coloured light was shone onto the sensor, the value on the latch would change appropriately.

But consider robot C. It is clear that the program itself has no means of knowing if the sensor is operating properly and hence if value stored in the latch is an accurate representation of the light detected by the sensor; however the value in the latch is now not in any way contingent on the ambient light conditions that pertain. Nonetheless, as the software executed is unchanged, the supervenience thesis suggests that the phenomenal states generated by the program must be the same; robot C must continue to 'see' red

For robot D, data from the colour sensor is no longer stored in the latch; instead the engineer has designed the circuitry so that the latch always stores the value $FF; once again, the program code executed by the robot is unchanged. And again the supervenience thesis suggests that the phenomenal states experienced by the robot will remain same. However, if the control-program for robot D was compiled using an 'optimising compiler' then the subsequent conditional branch would be replaced by a non-conditional branch; demonstrating that non-entered

conditional state sequences can be completely removed and the putative phenomenal states of the program must be unchanged, hence Chrisley's objection is invalid[128] and the original reductio holds.

Objections: (2) Computational states are not observer-relative but are intrinsic properties of any genuine computational system[129]

In addressing this objection I will initially consider the most primitive of computational systems - a simple two input / single output logic gate [X], with physical behaviour fully specified by the following table of voltage levels:

INPUT-1	**INPUT-2**	**OUTPUT**
0v	0v	0v
0v	5v	0v
5v	0v	0v
5v	5v	5v

It is clear that under **MAPPING-A**, (*+5v = COMPUTATIONAL STATE TRUE and 0v = COMPUTATIONAL FALSE*), the gate [X] 'computes' the logical **AND** function.

Conversely, under **MAPPING-B**, (*0v = COMPUTATIONAL STATE TRUE and +5v = COMPUTATIONAL FALSE*), it is clear that the gate [X] computes the logical **OR** function.

It follows that, at a fundamental level in the physical realisation of any logical system, such 'observer-relativity' must hold: the computational function of the system must be contingent on the 'observer-determined' mapping used[130].

---

128 Clearly, if the phenomenal experience of robot D differed from robot A, then the putative phenomenal states of a robot will always be contingent upon the particular type of compiler used by the roboteer (not on the semantics of actual program he or she wants to compile).

129 Objection raised by a member of the audience at the presentation of this paper at the 2006 'Computers and Philosophy' conference, Leval, France.

130 Although it is true that as the complexity of the logical system increases, the number of consistent computational functions that can be assigned to it diminishes, it remains the case that its computational properties will always be relative to the threshold

Further, it is clear that even if the physical-to-computational mapping is known, the function of the system remains observer-relative; that is, "different answers grow from the concerns of different individuals"[131]. Consider (a) a chess playing computational machine used to control the position of chess pieces in a game against, say, a human opponent and (b) the same program being used to control the illumination of a strip of coloured lights - the two dimensional chess board being mapped to a one dimensional strip of lights where the colour of each light is contingent on the value (king, knight, pawn etc) of the piece mapped onto it - in an interactive art exhibition. It is clear that the *purpose of the computations* is contingent on their *social use*. In Heideggerian terms, computing machinery doesn't exist in the world until it is put to some use - an event of '*breaking-down*' - such as playing chess when it becomes part of the background of '*readiness-to-hand*' required in the act of playing a game of chess; or interactively controlling an array of lights when it becomes an interactive piece of art. For Winograd and Flores, we see that, "for different people, engaged in different activities, the existence of objects and properties emerges in different kinds of *breaking down*". In these terms it is meaningless to talk about the existence of the computational system without concomitant purposeful activity and associated 'breaking-down'.

Conclusion

In this paper I have attempted to demonstrate (a) that computation is always fundamentally observer-relative and that (b) non-entered counterfactual state sequences of the control program of a robot cannot affect its putative phenomenal experience. Thus - being wary of panpsychism - the reductio argument presented herein should be seen to suggest that computers really *cannot* feel.

---

logic value used. The 'physical-state' $\Rightarrow$ 'computational-state' mapping will always co-determine the 'logical-function' that the physical computational system instantiates.

131 Cf. What is a word-processor?, in Winograd, T. & Flores, F. Understanding Computers and Cognition, Addison Wesley, 1986.

References:

Bishop, J.M., (2002a), Dancing with Pixies, in Preston, J. & Bishop, J.M., (eds), Views into the Chinese Room, (Oxford: Oxford University Press).

Bishop, J.M., (2002b), Counterfactuals Cannot Count: a rejoinder to David Chalmers, Consciousness & Cognition, 11(4), pp: 642-652.

Chalmers, D.J., (1996), Does a Rock Implement Every Finite-State Automaton?, Synthese, 108, pp.309-333.

Chrisley R., (2006), Counterfactual computational vehicles of consciousness, Toward a Science of Consciousness 2006, April 4-8, Tucson Convention Center, Tucson Arizona, USA.

Klein, C., (2004), Maudlin on Computation, (working paper).

Maudlin T, (1989), Computation and Consciousness, Journal of Philosophy 86, pp. 407-432.

Putnam, H., (1988), Representation & Reality, (Cambridge MA: Bradford Books).

Searle, J., (1990), Is the Brain a Digital Computer?, Proceedings of the American Philosophical Association, vol. 64, pp.21-37.

FORMAL ANALYSIS OF DYNAMICS WITHIN PHILOSOPHY OF MIND BY COMPUTER SIMULATION

Tibor Bosse, Martijn C. Schut, and Jan Treur
Department of Artificial Intelligence, Vrije Universiteit Amsterdam
http://www.few.vu.nl/~{tbosse, schut, treur}
{tbosse, schut, treur}@cs.vu.nl

Abstract. Computer simulations can be useful tools to support philosophers in validating their theories, especially when these theories concern phenomena showing nontrivial dynamics. Such theories are usually informal, whilst for computer simulation a formally described model is needed. In this paper, a methodology is proposed to gradually formalise philosophical theories in terms of logically formalised dynamic properties. One outcome of this process is an executable logic-based temporal specification, which within a dedicated software environment can be used as a simulation model to perform simulations. This specification provides a logical formalisation at the lowest aggregation level of the basic mechanisms underlying a process. In addition, dynamic properties at a higher aggregation level that may emerge from the mechanisms specified by the lower level properties, can be specified. Software tools are available to support specification, and to automatically check such higher level properties against the lower level properties and against generated simulation traces. As an illustration, three case studies are discussed showing successful applications of the approach to formalise and analyse, among others, Clark's theory on extended mind, Damasio's theory on core consciousness, and Dennett's perspective on intertemporal decision making and altruism.

Introduction

This paper introduces ideas on a research methodology that aims to bridge the gap between philosophical thought experiments and computer simulation. Computer simulations can be used as *intuition pumps*, to scale up and analyse such non-empirical experiments, as described by Dennett (2003):

> "Computer simulations … add further discipline: a way of discovering hidden assumptions of one's models, and a way of exploring the dynamic effects, by "turning the knobs" to see the effect of different settings of the variables. It is important to recognize that these computer simulations are actually philosophical thought experiments, intuition pumps, not empirical experiments. … Philosophers used to have to conduct their thought experiments by hand, one at a time. Now they can conduct thousands of variations in an hour …" (Dennett, 2003, Ch. 7, p. 218)

Even more so, the computer can be considered a tool to support philosophers in their thinking process on particular consequences of considered models:

> "… the evolutionary perspective … permits us to explore the interactions over time between agents that philosophers typically just handwave about. For instance, philosophers often ask "What if everybody did it?" as a rhetorical question, and don't stop to consider the answer, which they typically think is obvious. They never even address the more interesting question: What if *some* people did it? (What percentage, over what time period, under what conditions?)" (Dennett, 2003, Ch. 7, pp. 217-218)

However, setting up a simulation model often requires quite some technical work on programming in some programming language, which makes it a not very attractive activity for the average philosopher. Sometimes cooperation with a computer scientist or AI researcher who is interested in philosophical themes may help, but this type of researcher is also a bit rare.

The approach proposed here aims at improving this situation by offering computer-supported methods to obtain specifications that can be used within a computer at a conceptual modelling level, based on a gradual formalisation process of dynamic properties from a temporal linguistic and logical perspective. On the one hand this formalisation process provides a conceptual level specification of a simulation model describing the basic mechanisms underlying a process at the lowest aggregation level. Within a dedicated software environment that has been developed, this specification can be used to obtain a simulation model to perform simulations. On the other hand, conceptual level specifications of dynamic properties at a higher aggregation level of the process that is considered (for example, properties that are expected to emerge from the basic mechanisms) can be expressed, and formalised. Using available software tools, automated verification can be performed of such higher level properties against (1) the lower level properties of the simulation model, (2) generated simulation traces, or (3) empirically available traces.

The methodology has been successfully applied in case studies addressing major themes within Philosophy of Mind, in particular themes that involve phenomena with nontrivial dynamics. Philosophical themes that have been clarified in this manner concern dynamically emerging properties (such as representation relations; e.g., Jacob, 1997; Kim,

1996) for an overall process based on given or assumed mechanisms. In this paper, three case studies are described in which such emergent dynamic properties have been analysed. These case studies address the following themes: the (shared) extended mind (Clark, 1997, 2001; Clark and Chalmers, 1998) within ant colonies, the notion of core consciousness (Damasio, 2000), and the idea of altruistic behaviour based on intertemporal decision making (Dennett, 2003).

In Section 2 the conceptual analysis method involving specification from informal to formal format is discussed. Section 3 presents a first case study as an application of this method, on the use of extended mind (Clark, 1997, 2001; Clark and Chalmers, 1998) within ant colonies. In Section 4 it is shown how the method was used to analyse Damasio (2000)'s theory on core consciousness. Section 5 addresses Dennett (2003)'s perspective on how cognitive capabilities for intertemporal decision making play a role in the evolution of altruistic behaviour. Finally, Section 6 is a discussion.

Conceptual Analysis From Informal to Formal

Within our approach a *dynamic property* is considered to be a building block from which we can construct complex dynamic structures (cognitive agents, organisations, societies, complex systems). Typically, a dynamic property identifies a relation between something that happens at some time and something that happens at, possibly but not necessarily, another time. We distinguish a number of different formats for expressing dynamic properties. Depending on the contexts, these formats can be based on *informal natural language*, *semi-formal structured natural language* or a *formal language*.

An example of a simple dynamic property (in informal format) is the following: "If an agent observes that it is raining, then later on it will believe that it is raining". For less formal discussions, for example, with domain experts, informal (and semi-formal) formats are more appropriate than a formal format. On the other hand, if automated checking software is used, dynamic properties have to be in a formal format. A natural process during the analysis of philosophical questions is that first informal specifications (of dynamic properties) are expressed, and later these informal expressions are translated into semi-formal, and possibly into formal formats. A first step in such an analysis involves **acquisition of a (domain) ontology**. In this process, different **state properties** are identified and distinguished from each other: concepts that relate to an agent's **input state**, **output state**,

internal state, and to **external world states**. An example of an internal state property is belief(agent,_A itsraining). The ontology later facilitates the formalisation of dynamic properties, as the different concepts are already defined. Moreover, a formalisation of a scenario (for example empirical or imagined data over time) can be made by using the formal ontologies for the different states, in order to formalise a sequence of events as a temporal **trace**.

Usually, also a **temporal structure** has to be reflected in the representation of a dynamic property. This entails that terms such as 'at any point in time', 'at an earlier point in time', 'for all time points between t1 and t2', 'after', 'before' are used to clarify the temporal relationships between different fragments in the dynamic property. This temporal structure is a main aspect distinguishing informal properties from semi-formal properties. For example, the semi-formal variant of the informal property shown above is the following: "If at any point in time t1 and agent A observes that it is raining, then there exists a time point t2 after t1 such that at t2 the agent A believes that it is raining".

To obtain formal representations of dynamic properties, the input, output, internal and external ontologies are chosen as formal ontologies for states, specified by sorts, constants, functions and predicates within an order-sorted predicate logic language. In addition, the temporal structure, as present in a semi-formal representation, is expressed in a predicate logic format, using an ordering relation between time points, the usual logical connectives ($\wedge,\vee,\neg,\Rightarrow$) and universal and existential quantifiers over time ($\forall,\exists$). In our methodology, we use the temporal language TTL (Bosse et al., 2006a) for this purpose. In TTL, the property shown above is formalised as follows:

$$\forall t1\ \ [\ \text{state}(\gamma, t1) \models \text{observes}(\text{agent\_A}, \text{itsraining}) \ \Rightarrow$$
$$\exists t2 \geq t1\ \ \text{state}(\gamma, t2) \models \text{belief}(\text{agent\_A}, \text{itsraining})\ \ ]$$

Here, γ is a variable that stands for an arbitrary trace, and t1 and t2 stand for time points. Moreover, $\text{state}(\gamma, t) \models p$ denotes that state property p is true in the state of trace γ at time point t.

Using the formal ontologies, and the formalisation of the temporal structure, a formalisation is obtained of dynamic properties. This formalisation can be used to perform automated analysis and simulation within a software environment that has been developed for these purposes.

Dynamic properties can be specified for different **aggregation levels**, from the lowest level of the direct causal relationships between state properties within a process (modelling the basic mechanisms assumed) to higher aggregation levels for properties of the process as a whole

(modelling properties that emerge from the basic mechanisms). Within analysis, the different aggregation levels provide automated verification possibilities to check whether the higher level properties are consistent with, or even are entailed by the lower level properties. The properties at the lowest aggregation level are often specified in **executable** format, close to the format of a transition system or a finite automaton. This format is suitable as a basis for a simulation model, to obtain simulated traces of the process. In our methodology, the executable language LEADSTO (Bosse et al., 2005a), which is a sub-language of TTL, is used to specify such executable properties. The basic building blocks of this language are expression of the format $\alpha \twoheadrightarrow \beta$ (pronounced α *leads to* β), which informally means the following: if state property α holds for a certain time interval, then after some delay, state property β will hold for a certain time interval. For a precise definition, see (Bosse et al., 2005a).

Extended Mind (Clark)

As a first case study, using the approach described above, it is discussed how Clark's theory on extended mind was analysed and formalised (Bosse et al, 2005b, 2006b). This theory expresses that behaviour is often not only supported by internal mental structures and cognitive processes, but also by processes based on patterns created in the external environment that serve as external mental structures; cf. (Clark, 1997, 2001; Clark and Chalmers, 1998; Dennett, 1996). In particular, in the context of an ant society, where pheromone levels in the environment play a role as external mental state properties, the focus mainly was on (1) logical specification of a simulation model for the lower level mechanisms, and (2) at a higher aggregation level on the representational content (e.g., Bickhard, 1993; Jacob, 1997; Kim, 1996) of external mental state properties (i.e., the pheromone levels in the environment). The latter properties describe representational relations in a formalised form, for which it is to be verified whether they emerge in the process shown in simulated traces. Notice that in the case of an ant colony, the external pheromone states are used in a collective manner, they are shared by multiple agents.
Through modelling the following challenging issues on cognitive modelling and representational content were encountered in this case study: (1) how to define representational content for an external mental state property; (2) how to handle decay of a mental state property; (3) how can joint creation (by multiple agents) of a shared mental state

property be modelled; (4) what is an appropriate notion of collective representational content of a shared external mental state property; and (5) how can representational content be defined in a case where a behavioural choice depends on a number of mental state properties.
To model the ant society, the following ontology was used:

	body positions in world:
pheromone level at edge e is i ant a is at location l coming from e ant a is at edge e to l2 coming from location l1	pheromones_at(e, i) is_at_location_from(a, l, e) is_at_edge_from_to(a, e, l1, l2)
	world state properties:
edge e connects location l1 and l2 location 1 is the nest location location 1 is the food location	connected_to_via(l1, l2, e) nest_location(l) food_location(l)
	input state properties:
ant a observes that it is at location l coming from edge e ant a observes that it is at edge e to l2 coming from location l1 ant a observes that edge e has pheromone level i	observes(a, is_at_location_from(l, e)) observes(a, is_at_edge_from_to(e, l1, l2)) observes(a, pheromones_at(e, i))
	output state properties:
ant a initiates the action to go to edge e to l2 coming from location l1 ant a initiates the action to go to location l coming from edge e ant a initiates the action to drop pheromones at edge e coming from location l ant a initiates the action to pick up food ant a initiates the action to drop food	to_be_performed(a, go_to_edge_from_to(e, l1, l2)) to_be_performed(a, go_to_location_from(l, e)) to_be_performed(a, drop_pheromones_at_edge_from(e, l)) to_be_performed(a, pick_up_food) to_be_performed(a, drop_food)

An example of a semiformal representation and a formalisation of a dynamic property in the executable LEADSTO format (Bosse et al., 2005a) is the following (note that LP stands for 'Local Property', to be able to distinguish between Local (or executable, at lower aggregation level) and Global Properties (GPs, at a higher aggregation level)).

LP5b (Selection of Edge)

If an ant observes that it is at location A, and edge e1 connected to location A has the highest number of pheromones, compared to edge e2 connected to location A, then the ant goes to edge e1.
Formal representation:

observes(a, is_at_location_from(A, e0)) and connected_to_via(A, l1, e1) and
observes(a, pheromones_at(e1, i1)) and connected_to_via(A, l2, e2) and
observes(a, pheromones_at(e2, i2)) and i1 > i2

$\twoheadrightarrow$ to_be_performed(a, go_to_edge_from_to(e1, A, l1))

This is one of the executable dynamic properties that make up the logical specification of the simulation model that was used to perform simulations. For the complete specification of this simulation model, see (Bosse et al., 2005b).

The executable dynamic properties discussed above address the process at the lowest aggregation level (the local dynamic properties). The remainder of this section discusses dynamic properties of a higher aggregation level (in the TTL format by Bosse et al., 2006a) and their verification against lower level properties. Within these properties, γ is a variable that stands for an arbitrary trace. First a language abstraction is given:

food_delivered_by(γ, t, a) $\equiv$
  $\exists$l, e [state(γ,t) |= is_at_location_from(a, l, e)) &
  state(γ,t) |= nest_location(l) & state(γ,t) |= to_be_performed(a, drop_food)]

One of the properties considered at the highest aggregation level is:

GP1 Food Delivery Succesfulness
There is at least one ant that brings food back to the nest.
  $\exists$t $\exists$a: food_delivered_by(γ, t, a).

Another type of dynamic property at a higher aggregation level is a representation relation (e.g., Bickhard, 1993; Jacob, 1997, Kim, 1996, pp. 184-210) for the pheromone states of the environment. The backward case of a representation relation for the pheromone states in the environment involves a summation over multiple agents at different time points, and decay rate r with 0 < r < 1 is used to indicate that after each time unit only a fraction r is left; see (Bosse et al., 2006b):

Backward Representation Relation for Pheromone States
There is an amount v of pheromone at edge e, if and only if there is a history such that at time point 0 there was ph(0, e) pheromone at e, and for each time point k from 0 to t a number dr(k, e) of ants were present at e, and

$$v = ph(0, e) * r^{t} + \Sigma_{k=0}^{t}\, dr(t\text{-}k, e) * r^{k}$$

A formalisation of this property in the logical language TTL is as follows:

$\forall$t $\forall$e $\forall$v state(γ, t) |= pheromones_at(e, v) $\Leftrightarrow$
  $\Sigma_{k=0}^{t}\ \Sigma_{a=ant1}^{ants}$ case([$\exists$l,l1 state(γ, k) |= is_at_edge_from_to(a, e, l, l1)], 1, 0) * r^{t-k} = v

Here for any formula f, the expression case(f, v1, v2) indicates the value v1 if f is true, and v2 otherwise.

Likewise, according to the relational specification approach the following forward representation relation was specified.

Forward Representation Relation for Pheromone States

If at time t1 the amount of pheromone at edge e1 (connected to location 1) is maximal with respect to the amount of pheromone at all other edges connected to that location l, except the edge that brought the ant to the location,
then, if an ant is at that location l at time t1,
  then the next edge the ant will be at some time t2 > t1 is e1.

If at time t1 an ant is at location 1 and

for every ant arriving at that location 1 at time t1, the next edge it will be at some time t2 > t1 is e1,

then the amount of pheromone at edge e1 is maximal with respect to the amount of pheromone at all other edges connected to that location l, except the edge that brought the ant to the location.

A formalisation of this property is as follows.

```
∀t1,l,l1,e1,e2,i1
  [ e1≠e2 &
  state(γ, t1) |= connected_to_via(l, l1, e1) &
  state(γ, t1) |= pheromones_at(e1, i1) &
  [∀l2≠l1, e3≠e2 [ state(γ, t1) |= connected_to_via(l, l2, e3) ⇒
                 ∃i2 [0≤i2<i1 & state(γ, t1) |= pheromones_at(e3, i2) ] ]
  ⇒  ∀a [ state(γ, t1) |= is_at_location_from(a, l, e2)  ⇒
                 ∃t2>t1 state(γ, t2) |= is_at_edge_from_to(a, e1, l, l1) &
                 [∀t3 t1<t3<t2 ⇒ is_at_location_from(a, l, e2) ] ] ] ]
```

```
∀t1, l,l1,e1,e2
  [e1≠e2 &
  state(γ, t1) |= connected_to_via(l, l1, e1) &
  ∃a   state(γ, t1) |= is_at_location_from(a, l, e2)  &
  ∀a [ state(γ, t1) |= is_at_location_from(a, l, e2)    ⇒
                 ∃t2>t1 state(γ, t2) |= is_at_edge_from_to(a, e1, l, l1) &
                 [∀t3 t1<t3<t2 ⇒ is_at_location_from(a, l, e2) ] ] ]
  ⇒  ∃i1 [ state(γ, t1) |= pheromones_at(e1, i1) &
                 [∀l2≠l1, e3≠e2 [ state(γ, t1) |= connected_to_via(l, l2, e3)
                 ⇒ ∃i2 [0≤i2≤i1 & state(γ, t1) |= pheromones_at(e3, i2) ]]]]
```

The properties at a higher aggregation level discussed above and a number of other properties have been formalised and using a checking software environment have been (automatically) verified in simulation traces. This is a first manner for verification. A second way of verification is to establish logical relationships between properties (by mathematical proof). This also has been performed in a number of cases, under a number of assumptions. For more details, see (Bosse et al., 2005b, 2006b). The results of these verifications show that indeed in the process of the ant colony, for which the mechanisms are modelled at the lower aggregation level of the simulation model, the assumed representation relations for the external pheromone states emerge, which shows that these external pheromone states play the role of (collective) external mental states in the expected manner.

Core Consciousness (Damasio)

As another case study, Damasio's theory on core consciousness was analysed and formalised (Bosse et al, 2006c). According to this theory, a state of core consciousness (or conscious feeling) for a certain object occurs when an agent monitors a change of its representation of its body state (the protoself) after the occurrence of this object. In other words, the state of core consciousness represents the process of change of the

agent's body state representations co-occurring with the occurrence of the object, i.e., it represents transitions between the following states: *protoself at the inaugural instant - object comes into sensory representation - protoself as modified by the object* (Damasio, 2000, p. 177-178).

Based on Damasio's theory, first a formal model was provided of the states and basic processes leading to core consciousness. The building blocks of this model are state properties and their functional roles expressed by executable properties. The following ontology of state properties is used (describing a specific case study about an agent that listens to some very special music, and eventually becomes conscious about this music):

music	a beautiful piece of music is played
sensor_state(music)	the agent is perceiving the music
sr(music)	an internal sensory representation for the music is present
p	the agent's body is preparing to respond to the music
S	the agent is in a body state in responding to the music (e.g., by shivers)
sensor_state(S)	the agent is perceiving its bodily response S
sr(S)	an internal sensory representation for S is present
s0	the agent is in (initial) mental state 0
s1	the agent is in mental state 1
s2	the agent is in mental state 2
speak_about(music)	the agent speaks about the music

In addition, the following executable properties were identified to describe the basic mechanisms of the process at the lowest aggregation level considered.

LP0	music $\twoheadrightarrow$ sensor_state(music)
LP1	sensor_state(music) $\twoheadrightarrow$ sr(music)
LP2	sr(music) $\twoheadrightarrow$ p
LP3	p $\twoheadrightarrow$ S
LP4	S $\twoheadrightarrow$ sensor_state(S)
LP5	sensor_state(S) $\twoheadrightarrow$ sr(S)
LP6	not sr(music) and not sr(S) $\twoheadrightarrow$ s0
LP7	sr(music) and not sr(S) and s0 $\twoheadrightarrow$ s1
LP8	sr(music) and sr(S) and s1 $\twoheadrightarrow$ s2
LP9	s2 $\twoheadrightarrow$ speak_about(music)

Based on these executable properties simulations were performed.

The (backward) representation relation for the mental state for core consciousness s2 was specified as follows: ‘if no body state S and no music occur, and later music occurs and still no body state S occurs, and later music occurs and S occurs, then still later s2 will occur,’ and conversely. Formally:

Backward Representation Relation for Core Consciousness States

$$\begin{array}{l} \forall t1, t2, t3\ [\ t1 \leq t2 \leq t3\ \& \\ \quad state(\gamma, t1, EW) \models \neg S \wedge \neg music\ \& \\ \quad state(\gamma, t2, EW) \models \neg S \wedge music\ \& \\ \quad state(\gamma, t3, EW) \models S \wedge music \Rightarrow \\ \qquad \exists t4 \geq t3\ state(\gamma, t4, internal) \models s2\ ] \end{array}$$

$$\begin{array}{l} \forall t4\ [\ state(\gamma, t4, internal) \models s2 \Rightarrow \\ \quad \exists t1, t2, t3\ \ t1 \leq t2 \leq t3 \leq t4\ \& \\ \quad state(\gamma, t1, EW) \models \neg S \wedge \neg music\ \& \\ \quad state(\gamma, t2, EW) \models \neg S \wedge music\ \& \\ \quad state(\gamma, t3, EW) \models S \wedge music] \end{array}$$

This corresponds to the transitions indicated by Damasio (2000): *the proto-self exists at the inaugural instant - an object comes into sensory representation - the proto-self has become modified by the object*. For an alternative formalisation, based on the notion of second-order representation, see (Bosse et al., 2006c).

Similarly, when looking forward, the representational content of a mental state can be described by relating it to future world states. The future representational content of state property s2 can be informally described as follows: ‘if s2 occurs, then later the agent will speak about the music’, and conversely. In the logical language TTL, the expression is formalized as follows.

Forward Representation Relation for Core Consciousness States

$$\begin{array}{l} \forall t1\ \ [\ state(\gamma, t1, internal) \models s2 \Rightarrow \exists t2 \geq t1\ \ state(\gamma, t2, EW) \models speak\_about(music)\ ] \\ \forall t2\ \ [state(\gamma, t2, EW) \models speak\_about(music) \Rightarrow \exists t1 \leq t2\ \ state(\gamma, t1, internal) \models= s2] \end{array}$$

The backward and forward representation relations are dynamic properties at a higher aggregation level. Part of the analysis has been to automatically verify (using the SMV environment; cf. McMillan, 1993) that the lower level properties LP0 through LP9 together entail the representational content specifications. This confirms part of the claims made by Damasio (2000) in the sense that the suggested mechanisms as described at a lower aggregation level indeed entail the emergence of higher level dynamic properties that represent the process of monitoring how the agent’s body state is affected by a given object.

Intertemporal Decision Making and Altruism (Dennett)

The third case study is inspired by (Dennett, 2003 – chapter 7)'s discussion of altruistic behaviour from an evolutionary perspective; see also Sober and Wilson (1998), Trivers (1971). This case study concerns an analysis of how the occurrence of forms of altruistic behaviour within agent communities depends on cognitive capabilities of agents with respect to intertemporal decision making; see also (Darwin, 1871)[132]. The set up focuses on a population with *x* members which have some regular (weekly, monthly) interactions with each other. These interactions have the typical form that one agent provides something (a service) to another agent without immediate return. Examples of such interactions are lending money, or assisting the other agent with removal events. Each individual interacts with a subset of the population, which may not be the same set all the time. Each interaction has some future consequence. For example, I may be lending you money today (a 'giving' part of the interaction at time *t*), and after some time you will return me the money (a 'receiving' part of interaction at time *t'* > *t*). Based on the 'receiving' parts, individuals assign some *credit value* to individuals that they have had interactions with at different points in time.
Inter-temporal choice is a decision in which the realisation of outcomes may lie in the imminent or remote future. Recently, inter-temporal choice has caught the attention in the literature on behavioural decision-making [Loewenstein and Elster, 1992]. Before this, results on the subject were mainly due to the research contributions in related fields, like economics and animal psychology. The standard agent model for decision-making over time is a framework called time discounting [Loewenstein and Elster, 1992], which works according to the same principles as interest that one receives on a bank account: I calculate a delayed reward back to its current value based on the interest that I would receive for it.

We use a similar agent model for inter-temporal decision making here, extended to our particular decision situation (involving reciprocity for cooperation) in two main ways. Firstly, the decisions involve an explicit model the agent has of (regularities in) the environment, in this case incorporating the other agents. This results in parameters for *trust* of the agent in other agents. As explained below, the value of this parameter evolves over time as a consequence of monitoring (regularities in) the environment over time, i.e., the experienced (non)cooperations. Secondly, the individual decisions are concerned with choosing between (1) a possible reward in the remote future and (2) having no immediate cost, rather than choosing between an immediate and delayed reward (as

---

132 'As the reasoning powers and foresight of the members became improved, each man would soon learn that if he aided his fellow-men, he would commonly receive aid in return' (Darwin, 1871, p. 163).

investigated traditionally in time discounting). In the model, the discounted value $f_{discounted}$ of a future reward is calculated by: $f_{discounted} = f * 2^{-(1-\alpha)(t/n*(1-(tr+1)/2))}$, where

f : REAL = future reward,

$\alpha \in [0,1]$ = discount factor,

t : INTEGER = duration after which the future reward is received,

n : INTEGER = duration of cooperation, and

$tr \in [-1,1]$ = trust in the agent who asks you to cooperate.

If the discounted future reward evaluates higher than (or equal to) the current (immediate) cost, the agent decides to cooperate. In other words, *if $f_{discounted} \geq c$, then cooperate, else do not cooperate*, where c : REAL is the immediate cost.

It was tested how agents that use this decision function develop in a multi-agent society. The prediction was that these agents will show altruistic behaviour, will establish a larger social network than agents without such a decision function (i.e., agents that are not able to estimate the future reward, and thus never cooperate), and will eventually get a higher fitness.

Agents adjust their trust values in other agents according to the following principle: if I ask you to cooperate and you accept, then I increase my trust in you; if you decline, then I decrease my trust in you. For modeling such adaptation of trust over time, we use a trust function that was presented in [Jonker and Treur, 1999]. This function, as applied here, takes the response of an asked agent (accept/decline) to determine how to revise the trust value. Such a response $e \in [-1,1]$ evaluates to 1 if the agent accepts or -1 if the agent declines. A scaling factor $\delta \in [0,1]$ (which is constant throughout the experiments) determines how strongly an agent is committed to its trust values: a higher δ means that an agent puts much weight on its current trust value and lets an (non)cooperation experience not weigh so heavily; and vice versa. In the model, when the outcome of a request to cooperate is known, we calculate the trust value tr_{new} as follows:

$$tr_{new} = \delta * tr + (1 - \delta) * e$$

where

$tr \in [-1,1]$ current trust value,

$\delta \in [0,1]$ discount factor (constant),

$e \in [-1,1]$ the response of the agent who you asked to cooperate.

Thus, each agent maintains a list of trust values for all other agents in the environment. The model also includes a *cooperation threshold* $ct \in [-1,1]$ such that agent *x* only requests cooperation with agent *y* if trust of agent *x* in agent *y* is above this threshold.
The mathematical model for trust-based intertemporal decision making described above has been incorporated in a small number of dynamic properties that describe at the lowest aggregation level the basic mechanisms of the (societal) process considered here, thus providing an executable conceptual model for the simulation model. An example of a property that was specified is:

LP1 Trust Adaptation

Trust is adapted on the basis of experiences.

$\forall$x,y:agent $\forall$tr:real $\forall$e:real

has_trust_in(x, y, tr) $\wedge$ has_experience_with(x, y, e) $\twoheadrightarrow$ has_trust_in(x, y, delta $\times$ tr + (1 - delta) $\times$ e)

Here, 'delta' is a constant, e.g. 0.9. State property has_trust_in(x, y, tr) represents the fact that agent x has trust in agent y with value tr, and state property has_experience_with(x, y, e) represents the fact that agent x has an experience with agent y with response e.

Based on these properties at the lowest aggregation level a number of simulations have been made. Moreover, a number of dynamic properties at a higher aggregation level have been identified that are relevant for the domain of trust-based inter-temporal decision making. These properties have been formalised. Three of them are shown below (in an informal format).

FM Fitness Monotonicity

If x has the cognitive system for decision making, then there exists a time t such that for all t1 and t2 after t, with t1 < t2, the fitness of x at t2 is higher than the fitness of x at t1.

DMAF Decision Making Agents get Fitter

Eventually, all agents with the cognitive system for decision making, will be more healthy than the agents without this system.

NDMA Network of Decision Making Agents

All agents with the cognitive system for decision making will always cooperate with each other.

The above properties have been automatically checked against generated simulation traces involving up to 25 agents. They all turned out to hold, which validates the above statements, such as ‘decision making agents get fitter’, for the simulation traces. These results support the claims about the evolutionary survival value of a cognitive system for intertemporal decision making, as discussed by Dennett (2003, Ch. 7).

Discussion

This paper describes a computer-supported method to transform philosophical thought experiments into computer simulation, thereby ‘pumping up’ the intuitions of philosophers. The method involves both informal and formal conceptual analysis including specification of dynamic properties from linguistic, informal, through structured, semiformal, to formal, temporal logical formats. Within the developed software environment a dedicated editor is available to support this process from informal to formal specification. These specifications can be made both at the lower aggregation level of the basic mechanisms underlying the considered process and at the higher aggregation levels of dynamic properties expected to emerge from these basic mechanisms. Within the software environment, the former specifications can be used to perform simulation, whereas the latter type of properties can be checked automatically against simulated (or empirical) traces. Moreover, a model checker environment such as SMV (McMillan, 1993) can be used to verify whether these higher level properties are entailed by the lower level properties. The method is particularly relevant for those philosophical thought experiments where dynamics play a crucial role.

The method was illustrated by three case studies that were undertaken. For these case studies, dynamic properties at the lowest aggregation level of basic mechanisms were specified, constituting a simulation model, as well as properties at a higher level of aggregation, that are expected to emerge from the lower level properties. In each of the cases, an analysis was performed based on computer-supported formalisation, simulation and verification. These analyses supported claims made by Clark (1997, 2001), Clark and Chalmers (1998), Damasio (2000), and Dennett (2003), respectively.

Acknowledgements

The authors are grateful to Catholijn Jonker for many stimulating discussions about the methodology, and to David Wendt for his contribution to the development of the model described in Section entitled “Intertemporal Decision Making and Altruism”.

References

Ainslie, G. (2001). *Breakdown of Will*. Cambridge University Press, 2001.

Bateson, M., & Kacelnik, A. (1997). Starlings preferences for predictable and unpredictable delays to food. Anim. Behav. 53, 1129-1142.

Bickhard, M.H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 5, pp. 285-333.

Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A., and Treur, J. (2006a). A Temporal Trace Language for the Formal Analysis of Dynamic Properties. Technical Report, Vrije Universiteit Amsterdam. http://www.few.vu.nl/~treur/TTL.pdf

Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J. (2005a). LEADSTO: a Language and Environment for Analysis of Dynamics by SimulaTiOn. In: Eymann, T. et al. (eds.), *Proc. of the 3rd German Conference on Multi-Agent System Technologies, MATES'05*. LNAI 3550. Springer Verlag, 2005, pp. 165-178.

Bosse, T. Jonker, C.M., Schut, M.C. and Treur, J. (2005b). Simulation and Analysis of Shared Extended Mind, *Simulation Journal: Transactions of the Society for Modeling and Simulation International*, vol. 81, 2005, pp. 719 - 732.

Bosse, T., Jonker, C.M., Schut, M.C., and Treur, J. (2006b). Collective Representational Content for Shared Extended Mind. *Cognitive Systems Research Journal*, vol. 7, 2006, pp. 151-174.

Bosse, T., Jonker, C.M., and Treur, J. (2006c). Formal Analysis of Damasio's Theory on Core Consciousness. In: Fum, D., Del Missier, F., Stocco, A. (eds.), *Proc. of the 7th International Conference on Cognitive Modelling, ICCM'06*, 2006, pp. 68-73.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. MIT Press, 1997.

Clark, A. (2001). Reasons, Robots and the Extended Mind. In: *Mind & Language*, vol. 16, 2001, pp. 121-145.

Clark, A., and Chalmers, D. (1998). The Extended Mind. In: *Analysis,* vol. 58, 1998, pp. 7-19.

Damasio, A. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. MIT Press.

Darwin, C. (1871). *The Descent of Man*. John Murray, London.

Dennett, D.C. (1996). Kinds of Mind: Towards an Understanding of Consciousness, New York: Basic Books.

Dennett, D.C. (2003). *Freedom Evolves*, New York: Viking Penguin.

Jacob, P. (1997). *What Minds Can Do: Intentionality in a Non-Intentional World*. Cambridge University Press, Cambridge.

Jonker, C.M., and Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust based on Experiences. In: F.J. Garijo, M. Boman (eds.), *Multi-Agent System Engineering, Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99*. Lecture Notes in AI, vol. 1647, Springer Verlag, Berlin, 1999, pp. 221-232. Extended version as Technical Report.

Kim, J. (1996). *Philosophy of Mind*. Westview Press.

Loewenstein, G.F., and Elster, J. (1992). *Choice over time*. Russel Sage Foundation, New York, 1992.

McMillan, K.L. (1993). Symbolic Model Checking: An Approach to the State Explosion Problem. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1992. Published by Kluwer Academic Publishers, 1993.

Sober, E., and Wilson, D.S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behaviour*. Harvard University Press, Cambridge, MA, 1998.
Trivers, R.L. (1971). The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*. 46:35-57, 1971.

DUALISM AND MATERIALISM, INCOMPATIBLE?

GERHARD J. DALENOORT
Groningen,

The Netherlands

g.j.dalenoort at xs4all.nl

Former affiliation: University of Groningen, The Netherlands.

Mind and brain are traditionally considered as two different systems, of which it must be accepted that they are related, but that belong to very different categories. To avoid the dilemma's of the so-called mind-body problem, the only possibility is to consider the brain as a system that can produce mind-like properties. To solve the various subproblems of the mind-body problem, the way physics has solved comparable problems at the atomic level (1900-1930) can show the way cognitive science can cope with the mind-body problem. The outcome is that a multiplicity of models is needed, that seemingly are incompatible. The paper argues that the seeming inconsistencies are due to our preconceived notions, and that they can be solved by resetting these notions. This involves the construction of correspondences between the models. The approach finally is aplied to the problem of specific energies of J. Müller.

1. Introduction

From the existence of an abundant literature on the mind-body problem, starting from the underlying question 'materialism or dualism?'(exclusive or), one might conclude that it is generally assumed that dualism and materialism are incompatible. As a consequence the mind-body problem does not seem to have a solution, and many related questions, such as 'Can machines think?', have been answered by a plethora of mutually incompatible answers. The question 'Can machines think?' falls under the topic of this conference, 'Computers and Philosophy', the same is true for such questions as to the necessity of consciousness and emotions for mental functioning. There are no convincing arguments that only the reductionist view is correct, nor that only the dualist position is correct. It also seems impossible that both can be correct, since each seems to exclude the other view, they are incompatible in their strong forms. I shall argue that the essential ideas can be "saved", but in a form that leaves room for the other view. The dilemma of incompatibility of the opposing views can be solved by giving up some preconceived notions that may be considered as to belong to folk-psychology. Physics can show in what way it has solved comparable problems in the domain of atomic processes (quantum mechanics), and problems handled by the special and the general theory of relativity. One notion that had to be abandoned was that a "thing" (or object) cannot simultaneously have particle-like properties and wave- like properties, another one that it is not possible to determine simultaneously the values of variables like position and velocity (Heisenberg uncertainty relation). Also the fundamental assumption that an accelerated charged body (like an electron) must emit radiation, had to be given up for the atomic domain (for an overview accessible for non-physicists, and with a keen eye on the epistemological problems, see Lakatos, 1970/1978). For cognitive science there is in addition another type of problem, the impossibility to measure consciousness, and qualia in general, in a non-subjective manner. On the other hand, individually we have direct experience of phenomena involving consciousness, and we can communicate about them by means of language. This in fact makes the problem from the cognitive point of view more accessible than the problem how an electron "behaves" inside a hydrogen atom. For cognitive science it is in the first place essential that the idea of 'mind and brain' is abandoned in so far as it may suggest that they constitute two different, interacting systems, that together can *produce* "mysterious" properties like mind, consciousness, and qualia, but also mental disorders (Note 1). It is of great historical interest that already Spinoza (round 1650) had come to that conclusion, a fact of which the significance has recently been emphasized by Damasio (2003), in order

to eliminate problems in the interpretation of emotions. Another indirect suggestion in this direction had been made by Pascal (1670/1962), in one of his Pensées: "To understand the parts (of a system) one must know the whole, and to understand the whole one must know the parts" (in the sequel called 'Pascal's dictum'). It clashes with the naive reductionist view, as well as with the view that the categories we use to describe mind are essentially different from, and incompatible with the categories we use to describe brain. The problem resides on the one hand in a naive view of what reductionism is, and on the other hand in strictly maintaining categories like matter and mind.

The approach proposed here allows to interpret in a "natural way" the bewildering variety of models and phenomena of mind and brain, from a large diversity of scientific domains. These domains are psychology and psychiatry, with their large sets of subdomains, language and linguistics, artificial intelligence on one hand, the neural sciences and neural networks on the other hand. Especially the phenomena of mental diseases and deviations from what we consider as "normal" pose problems, in the first place because scientists rarely have direct individual experience of their consequences, and secondly because such individual experience cannot easily be used scientifically. A central problem is the variety in methodology and reasoning between the various sciences mentioned. In the natural sciences concerned with the brain, and with respect to artificial nervous systems in general, there now exists a generally accepted unity as to the desired methodology and criteria for the validity of arguments. For the other sciences concerned with - what we call here - the "products of the brain", like psychology, linguistics, and the medical sciences as far they are concerned with mental disorders, there does not exist a unity in methodology and reasoning. A still different type of reasoning exists in the domains of various types of artificial neural networks (e.g. connectionism), and the so- called functionalist approach in psychology. The latter constructs models based on computer-like programs, like ACT-R and SOAR, and the same holds, in a different form for the approach of Artificial Intelligence. It aims at constructing computer programs (and robots) that simulate human behaviour, including mental behaviour, but without the aim that the underlying processes, architecture, or even logistics, are the same as in the biological system. These questions are part of the more general question: what type of model is appropriate to answer certain questions? Can a purely functional model of a cognitive phenomenon, without reference to the underlying neural processes, be the end of a research project? Will our understanding of that cognitive phenomenon be improved if we would know the underlying processes? According to Pascal's dictum it would. Understanding can only take shape in the form of a model; what type of model for the neural processes is appropriate?

Connectionism does not model the biological aspects properly, and also the structural and logistic basis of connectionist systems is far from that of the biological systems.

A fundamental distinction between the domains that are concerned with 'brain' on the one hand, and 'mind' on the other hand, is related to the fundamental distinction between a causal view of the world, and a goal-directed view. These views are mutually incompatible from a naive point of view, but provide a good opportunity to see how the approach proposed can serve to solve the dilemma. The dilemma is that we as subjects experience our intentions and purposeful behaviour as very real, whereas the natural sciences only leave room for a causal world. The solution is that by means of the phenomena of self-organisation we can show how causality is related to goal-directedness, where goal-directedness is the underlying necessary quality - at the biological level - for us to experience purposefulness. The latter then is an example of a "mysterious property", that does not fit into our common notions of daily life. In this context it is remarkable how in daily life and speech we can easily switch between causal and goal-directed terminology, in what at first sight may seem a chaotic manner. Closer investigation usually reveals there is some system in it. The distinction is also related to the distinction between constructed (programmed) systems and self-organizing systems, and as such relevant to the topic of a conference on computers and philosophy.

2. Epistemological Considerations

2.1. Lessons from Epistemological Problems in Physics

Physics was in the last decennia of the 19th, and the beginning of the 20th century, confronted with a number of epistemological problems. The way physics has dealt with these problems has a bearing on epistemology in general, notably on the role of our *a priori* notions. These concerned - among others - the possible or impossible architecture of the atom, in relation to the experimentally known line spectrum of hydrogen atoms, and the outcome of the Michelson-Morley experiment that the velocity of light is constant, irrespective of possible movement of the observer or of the source of the light (for an overview see Lakatos, 1970/1978, Section 3). At that time no physicist did, or

could surmise what would be the final result of the changes by, say, 1925. By then there was little discussion any more on the formal part of the theory (the mathematics), but there were very different opinions on the interpretation of the formulas. There are - at least - two lessons to be learned in general: the first is the role of *a priori* notions, the second the way we ascribe properties to objects and systems. These are relevant for the way we should face fundamental problems in cognitive science at present.

A priori notions are part of *a priori* models derived from our cognitive and perceptual experiences in daily life, especially our imagination. We must assume that the nature of our imagination is based on the (macroscopic) environment in which we grow up as individuals ("nurture"), but also on genetic factors determined by the evolution of man, and that in turn determine the way we think and make mental images. The prime example of fundamental physics concerns electrons, for which we cannot imagine that they can behave as particles as well as waves, depending on the conditions. According to our notions of daily life, a thing cannot behave as a particle at one moment, and as a wave at the next, or even show both types of behaviour at any moment, dependent on how we measure. Another example is our notion of causality, and the way it probably is related to psychological conditioning. It is a measure of the originality of David Hume that he realised long before the advent of quantum mechanics, and of Pavlov, that our notion of causality must be purely based on the steady occurrence of two events, always in the same sequential order in time. This is the reason that in folk psychology lightning is seen as the cause of the following thunder. The history of physics can provide guidelines to discover our *a priori* notions with respect to mind and brain, which in my opinion are at the roots of the so-called mind-body problem, although there are important differences between psychology and physics. The new model for electrons, and for other elementary particles, explicitly incorporated the possibility of a fundamental duality and complementarity of behaviours, dependent on the conditions of an experiment or observation. It had to be accepted that the electron could in no way be compared to any solid object as we encounter in daily life. The lesson for cognitive science is that the solution to the seeming incompatibility of materialism and dualism cannot be found if we stick to our notions of daily life and of "folk psychology". The only way out is to basically accept that the brain is a material system, that can produce consciousness, qualia, and so on. The fact that phenomena of consciousness *seem* to belong to a category that is entirely different from material processes, can not be an argument (Note 2).

The scientifically appropriate task of the sciences concerned with mind and brain is to discover under what conditions matter can produce consciousness. This is a specific case of a more general, but modified principle of scientific reduction (Dalenoort, 1987), which in turn is a case of constructing the correspondences between different levels of description. It must come in the place of naive notions on the possibility of scientific reduction.

2.2. Ascribing Properties

It is important to know that some questions are wrong questions, they simply have no meaning, such as asking 'What sound does the moon make?', and also why they are wrong questions. Another example, relevant for our topic, is the meaninglessness of the question 'What is gravity?'. It is equally unanswerable as the question 'What is consciousness?' The concept of gravity is a construct; it is part of a mental model; we cannot observe gravity nor measure it. We can only measure observables of *phenomena*, that we in this case *ascribe* to the force of gravity according to a model, preferably a general model; we do not want a specific model for every separate phenomenon. The importance of the model of Newton was that by means of the concept of gravity, he could construct a single formalism by which he could analyze and do calculations on such different phenomena as the falling apple and the structure of the solar system. The same formalism can also be used to calculate the orbits of satellites. It is equally senseless to ask 'What is consciousness'. Like in the case of gravity, we can only ask under what conditions certain phenomena will occur that we ascribe to consciousness. There are differences between gravity and consciousness in experience, but this is purely subjective. This experience cannot qualify as a measurement as we usually interpret this concept. Of course, the fact that we now have language, and can talk about consciousness and related phenomena, can provide a degree of intersubjectivity. We can even discuss whether language might be a condition for higher levels of consciousness, for example in relation to our interpretation of the nature of consciousness in dogs. This new approach has various consequences. One is the way we ascribe properties to systems. As already stated, in this context an old idea, probably first explicitly stated by Blaise Pascal (1660/1962), is relevant: "To understand the whole, one must know the parts, and to understand the parts, one must know the whole". Intuitively one can experience the relevance of Pascal's dictum. My favourite example is a cogwheel. Imagine that someone, perhaps you, sees a cogwheel for the first time in

his or her life, lying on the table. Would you know what it is for? Probably most boys will have their knowledge from the experience of taking an old alarm clock to pieces, which can show the way cogwheels work. One may also have seen old church clocks. Whatever is the case, one can only grasp the use of the cogwheel from somehow seeing it operate in a whole. On the other hand, if one wants to understand how an old mechanical clock works, one must go down to the single cogwheel, and see how the teeth subsequently push the teeth of the next cogwheel. The fact that they sit on the circumference of a wheel, makes the process can go on forever. Here the elementary sensation, and subsequently *notion*, of pushing is essential. Pascal's dictum leads to a direct reconsideration of the roles of the contexts of discovery and of confirmation. The example of the cogwheel shows that understanding, knowledge, meaning, originate from a complex process of suitable interactions, between an observer and the world. Once we have understood some part or aspect of the world, we may feel the need to formalise the acquired knowledge. But if we then go to epistemological questions, and forget about the way we acquired the knowledge, we may encounter problems. The formal picture is an abstraction, it is in itself a model, often a metamodel. Ontology is concerned with the nature of our knowledge and of the world, whereas epistemology is concerned with the question how we can acquire knowledge of the world, and what we can learn about the world in the first place. It then is essential to take into account the process of discovery, and to make aspects of our perception, and the actual experience of obtaining and having insight, part of our discussion on epistemological questions. It is my feeling that the mystery of consciousness and qualia is at least partly due to not realising that consciousness is a product of interactions of the parts of the brain, and interactions with the outside world through the senses and motor experiences. We can compare this to a familiar example: a conversation. It emerges from the interaction of two persons, and it can have a content and meaning that can also not be reduced to the sounds involved, or to the rules of grammar. This phenomenon is as complex as our conscious experience, and we hardly feel an existential problem with it, as many people seem to feel when they utter their surprise about us being conscious. The only epistemological stand is to accept the phenomenon, and to ask in what forms it can show up, and under what conditions. This is also the proper form of the reductionist quest (Dalenoort, 1987). It is one of a dualist pair, according to Pascal's dictum: we must fully realise that consciousness is a property of the system as a whole. It makes no sense to ask what the site of consciousness is in our brain. All actual properties can be considered as the products of interactions, between an observer and a system, or between two or more systems and an observer. Our talking about properties as always being there, independent from the observer, is an

abstraction. If we say 'ripe tomatoes are red', without actually seeing them, we refer to a situation that we will experience the colour of red when we see ripe tomatoes; we know this from previous experiences, and we would be surprised if someone told us upon seeing green tomatoes, that they are fully ripe. The actual experience only occurs if we are in interaction with the object. The statement 'ripe tomatoes are red' is a construct, which only makes sense by its implicit reference to a specific situation. The colour 'red' in this case is the product of an action of white light, that after having been reflected and modified in constitution by the tomato, enters into our visual system. (We here use the term 'interaction in a very general sense).) A hydrogen atom "exists" on the basis of the interaction between its nucleus, a proton, and its electron. A hydrogen atom can only be "seen" if it emits a photon, and the photon must hit our retina (one photon actually is not quite enough). We can generalise that all actually occurring properties are the products of interactions, and that statements on properties, for example as occurring in a formal model, are abstractions, the term 'property' then refers to a construct.

3. Two Science-Theoretical Problems

3.1. Consequences for our Fellow-Computers

A well-known scientist of Artificial Intelligence once wrote in a review of a book, that he did not mind to be compared to a computer - as a human in general - but strongly objected to the idea that it could be a simple computer (P. Hayes, about 1975). On the other hand, we may ask if a computer will ever be comparable to us, at some time in the future. If it would, we do have an idea of what the consequences could be, from a relatively large number of science-fiction books and films. From sophisticated analyses and discussions of the Turing test, and elaborations of it, we can conclude that there is no essential argument that we would be able to decide whether a machine that states 'I am conscious' would not be speaking the truth. A simple argument would be that we can also not conclude that a person - assuming we can see that a person is speaking and not a computer - speaks the truth when he or she says 'I am conscious'. As already noted, the basic argument is that we have no objective method to measure consciousness, like we can measure the size of an electric current. Our conclusion that someone must be conscious who speaks sensible words, is based on a model, or

a metamodel, of the world, in particular of human beings in that world. Such considerations are relevant if one wants to discuss the role of computers for our view of the world, of our philosophy (ontology and epistemology). What do people have that computers do not have? From what we have seen so far, there is no reason to believe that we cannot simulate human thinking, and that we cannot make computers creative. But we also know that there is, at least until now, an essential difference between humans and computers: the former are made of "flesh and blood". It might well be that the material constitution of a piece of matter that claims to be conscious, is an essential factor. Therefore, we can only continue to analyse the structure of our reasoning, and the architecture of our knowledge by means of computers, and attempt to discover the correspondences to neural processes. We can also have phantasies about the differences between the 'mental processes' and 'feelings' of humans and computers, but we must be aware of what is knowledge, and what belief.

3.2. The Problem of Specific Energies of Johannes Müller

Johannes Müller (1841) asked the question how a part of the brain could "know" what was the "meaning" of a given input, for example, whether something is coming from the eyes, or from the ears, whether it meant a tickle on one's leg, or a pain in one's hand. Müller's problem is also known as the problem of 'specific energies'. He surmised that the signals coming from different sense organs could be recognised by the type of energy involved. At that time there was not yet a clear conception of energy, nor of the properties of neurons, let alone of the existence of spikes. We could now give the tentative answer that the spike frequency of neural signals coming from the eyes is different from that from the ears, but there is no experimental evidence that this could be the case, nor does it solve the fundamental problem in the first place. Müller's question nowadays still has no good answer, but from the approach advocated in the sections above, we can reason that the question is ill-posed. Let us think a bit along the old lines first. We know that the first perceptual cortical areas in the brain are specific: the visual cortex only receives input from the eyes, the auditory cortex from the ears. But consider our ability to hear a dog bark, and to see it simultaneously, and realise that that is where the sound comes from. In a naive type of model we would still be confronted with the question how this is possible: there must be a part of the brain where the signals from both cortices come together, a necessary requirement to draw the conclusion. The problem would be shifted to a more central part of the brain (see also Dalenoort,

1996). William James (1890), half a century later, gave a meticulous analysis of many phenomena of perception, and possible explanations, but the problem of specific energies was not solved. I shall argue that the problem as phrased by Müller, and others, is a pseudoproblem. It is of the same type as Ryle's example of the visitor of Oxford who asks "Where is the University of Oxford?". It has no answer, since it is an abstract concept (Ryle, 1949). The "object" concerned is a functional, or collective property of a large number of "things" in Oxford. How is a similar problem solved in an organization of humans, say a firm, or a university department? For example, how do we know the origin of a message? It is common that a message has the name of a sender with it, and also in our electronic era we often know the sender, unless it is spam. But neurons cannot "read", or recognize the sender, they cannot 'know' at all, they can only receive spikes, and these add to, or deduce from the excitation already present in the neuron. The neuron can also not "know" where the spikes come from. So the problem remains: what makes an excitation in the brain specific, such that we can have the conscious experience of a certain specific meaning? Here we have to build upon a generally accepted hypothesis: for every mental phenomenon or experience there must be a corresponding neural process. What is the neural counterpart of our specific experience of that black dog barking on the other side of the street? To find a tentative solution, we can reason backwards. Even if the visual cortex receives input only from the more peripheral parts of the visual system, the problem remains, since the neurons everywhere can do nothing else than receive spikes, and produce spikes if their excitation level is above a certain threshold. Of all the parts of the visual system, the most specific part is the eye, especially if one considers colours. We must envisage the possibility that neural activity in the retina contributes to our specific conscious experience. This hypothesis is strengthened by the difference we experience between seeing a red tomato, and with eyes closed imagining it lying there on the table. It seems very likely that the activity of the retina is essential to explain the difference. We can generalise that all (conscious) experiences are the products of the interactions of all the parts taking part in the underlying excitation. That the experience occurs in a human being, and probably not in computers or robots, would then be due to the difference in the type of matter and processes underlying the experience; brains are biological systems. But as noted before, we cannot do conclusive measurements, and know for sure that the robot is not conscious. Similarly we may surmise that if we feel a pain in our left hand, that the neural activity in that hand is an essential part of the underlying process, notably with respect to the location of the pain. Also on this point a problem occurs if we would not accept this idea: the consequence of the experience being produced only by central neural activity in the brain, would lead to the question

how an excitation, again, of the pain in our left hand, could obtain the specificity of the location needed. We again stumble on Müller's dilemma (Note 3). We here then have an extension of the earlier picture on the way we should interpret properties in general: as the products of interactions, between systems, and of the system as a whole with an observer. The difference with an external observer is that the observer involved in conscious experiences is built-in. This observer does not exist independently from the system, he or she must be seen as a functional property of the system. The term 'collective property' is also appropriate, since it expresses that no single mental property - reaching the conscious level - resides in a single part of the brain. Analogous to the properties of the hydrogen atom that emerge from the interaction between the proton and an electron, the property 'consciousness' emerges from the interactions between all the parts of the brain and body. The term 'emergent property' has often be used to express the same idea, but here we emphasize that also physical properties are emergent. As a consequence we can obtain a general epistemology, where concepts do not only apply for the qualia of psychology and philosophy. We ascribe different types of consciousness to dogs, chimpanzees, and, say, spiders. The question then arises which are the essential aspects of nervous systems that lead to various forms of consciousness. The proper question to ask then is not how we can "reduce" a certain conscious experience to certain neural properties (a naive form of scientific reductionism), but which architecture, properties, and processes of a nervous system correspond to a certain conscious experience.

Notes

1. The notion of 'to produce' and of 'producer' originate from the American philosopher E.A. Singer Jr. (1959), Chapter 18. He introduced the term as a more neutral term than the terms of 'cause' and 'effect', which may have to explicit connotations for different people. Here it is used in the same vein: to have a neutral term available, without implications of a priori notions. Many of the points I have made, are also discussed by Singer, here I have not made an attempt to integrate his ideas with mine.

2. In the approach of General Systems Theory one might argue that mind and brain may also be considered as two subsystems of a single, larger system. There are no a priori boundaries of systems, these depend on the questions one wants to answer, and on the models chosen to describe the system; again to large extent a priori notions. But for mind and brain this is in fact not acceptable, because the two

subsystems belong to very different categories; splitting into subsystems is only sensible for real subsystems, of similar types.

3. Of course there is the problem to understand the phenomena of experienced pain in a phantom limb, or pain in a foot or leg as a consequence of a hernia, although there is no tissue damage in the foot or leg. One might call this a phantom pain; the most likely hypothesis for the cause is that damage at the place of the hernia, causes the "erroneous" excitation patterns in the nervous system that are characteristic for these pains, and that are not like a pain caused by a damage in the foot or leg. Notably the location of the pain is erroneous.

References

Dalenoort, G.J. (1987), Is physics reductionistic?, In: Wm. J. Baker, M.E. Hyland, H. van Rappard, and A.W. Staats (Eds.), *Current issues in theoretical psychology*, North Holland Publ., Amsterdam.

Dalenoort, G.J. (1990), Towards a general theory of representation, *Psychological Research* **52**, 229-237.

Dalenoort, G.J. (1996), Johannes Müller and the systems approach to the problem of specificity of the memory trace, *Proc. 15th Cheiron Conference*, Leiden, The Netherlands, 1996 (published by the Dept. of Psychology, Leiden University, The Netherlands), pp. 31-36.

Damasio, A. (2003), *Looking for Spinoza: Joy, sorrow, and the feeling brain*, Harcourt, Orlando, USA.

James, W. (1890), *The principles of psychology*, Vol. **2**, Henry Holt Co., (republished by Dover, 1950).

Lakatos, I. (1970/1978), Falsification and the methodology of scientific research programmes, in Lakatos, I. and Musgrave, A., *Criticism and the growth of knowledge*, Cambridge Univ. Press, 1970, reprinted in: Lakatos, I., *The methodology of scientific research programmes*, edited by J. Worrall and G. Currie, Cambridge Univ. Press. (1978), especially Section 3, p. 31-89.

Müller, J. (1834-1840), *Handbuch der Physiologie des Menschen*, Hoelscher, Coblenz.

Pascal, B. (1670/1962), *Pensées*, edited by J. Chevalier, Livre de Poche, Paris, 1962), Nr. 84, p. 56 (in other editions this number is 72, the cited text is towards the end of this 'thought', it has the title 'Disproportion de l'homme').

Ryle, G. (1949), *The concept of mind*, Hutchinson, London.

Singer, E.A. Jr. (1959), *Experience and reflection*, (Edited by C. West Churchmann), Univ. of Pennsylvania Press, Philadelphia, USA.

BLOCKING BLOCKHEAD

Mehdi Nasrin

Departmnet of Analytic Philosophy

Institute for Studies in Theoretical Physics and Mathematics

P.O.Box: 19395-5746

Niavaran Square, Tehran, IRAN

Phone: (+9821)2280-3669 Email: nasrin@ipm.ir

Introduction

Alan Turing, in his celebrated paper (Turing 1950) tackles a very crucial question in philosophy of mind: 'Can machines think?' He then argues that in order to answer this question one should begin with definitions of the meaning of the terms 'machine' and 'think'. Turing later in his paper explains what he means by 'machines'. Machines are physically implemented discrete state automata or what he calls "universal digital computers". But he is not clear about the notion of 'thought' and

(deliberately) avoids giving a definition of this notion. Turing introduces the imitation game in which a human judge tries to distinguish between a human and a machine just by asking them some questions. Turing tries to replace the abovementioned question with another question and argues that answering the second question is more fruitful and practical. The second question is 'How good are computers in an imitation game'? The main philosophical question, then, is to examine if we are justified to attribute thoughtfulness to a machine which is regularly successful in an imitation game (i.e. regularly passes the Turing test). In the first part of my paper, I present my understanding of Turing's claim and shall argue that, as Dennett has argued, one should take Turing's claim as a scientific claim which can be (dis)confirmed by actual facts.

In the second part of this paper, I shall examine the common ways of attacking the Turing test conception of intelligence (TTCI), according to which what we intuitively count as non-intelligent machines can frequently pass a typical Turing test. Ned Block introduces one of these machines (Blockhead) whose mechanism, in order to pass the test, is nothing but looking up in a huge database and randomly choosing a string. Blockhead restores all the possible answers that a particular human being might produce, therefore if the test is conducted between this machine and that human, the interrogator would not be able to unmask the machine. We have a very strong intuition that such Blockhead, which lacks the built-in capacities of learning, problem solving and such, is not thinking and therefore is not intelligent.

I shall provide two replies to Block's objection. According to a more particular one, Block's objection could be answered by showing that the non-intelligent Blockhead which was said to be able to pass the Turing test actually could not pass it (despite appearances). To this end, I shall argue that there are some strategies that the interrogator can use to check if the machine is just recalling saved information or also able to learn and improvise. So we shall see that Blockhead in its original design cannot pass the test and it should be armed with some other mechanisms in order to pass the test. However by adding more mechanisms to what Blockhead already has, we get distance from that original convincing intuition that it is not intelligent. Therefore, we cannot easily indicate that the new machine (with a wide variety of embedded mechanisms) is not an intelligent machine. I shall argue that Block has adopted a very restricted conception of the Turing test and if we take a broader conception, the one that Turing intended to propose, we see that their counter-examples fail. Also, I will argue that it is logically possible that the interrogator ask an infinite number of questions in a limited amount of time (inspired by George Boolos' Zeus machine). So the counter-examples can be answered by showing that the non-intelligent machines in these thought experiments cannot pass the test.

A more general reply shall be offered for all logically possible counter examples against the Turing test conception of intelligence. The test, as we will see, is not an operationalist definition of intelligence. Therefore, the existence of counter-examples should not be regarded as prima facie evidence against it. We have learned from Quine that no statement is immune to revision. One interpretation of Quine's sentence is that for any statement, there is a possible world such that that statement does not hold in that possible world. It is not "difficult" to provide logically possible counter-examples to any revisable claim. A computer which produces a string of characters randomly could pass the test out of sheer luck. However, this possibility does not create a serious problem for Turing. It seems thought experiments invoking nomologically impossible scenarios should not be considered as real counter-examples. Let us start with examining the TTCI.

The Elementary Game

Turing's project of finding a proper alternative for the central question 'Can machines think?' starts with a game (call it the Elementary Game) which is played by three people, a man, a woman, and an interrogator who may be of either sex. The interrogator does not see the others but can ask questions to them. "The object of the game for the interrogator is to determine which of the other two is the man and which is the woman" (Turing 1950, p. 434). One of the other players (be it the man or the woman), tries to help the interrogator in her task, while the other player tries to deceive the interrogator about the gender of the two players. Let us call the other players respectively "Pro-interrogator" and "Anti-interrogator", or "Pro" and "Anti" for short. The Elementary Game is based on the questions that the interrogator asks to Pro and Anti. Turing also makes some suggestions to prevent the interrogator from performing her duty by using non-linguistic abilities. He sets up the test in a way in which the irrelevant features, like tone of voice, are carefully screened off so that only essential features of intelligence, according to Turing, can be examined. Given these restrictions, the interrogator's only source will be her conversations with the other two players[133].

---

133 Dennett (1986, p. 5) indicates that these restrictions come from the same insight which underlies "the new practice among symphony orchestras of conducting auditions with an opaque screen between the jury and the musician". Since what matters for a musician is not features like sex, color, weight, hair style but "musicianship", the set up prevents the judge to observe these non-essential features and make a biased verdict.

The Elementary Game ends[134] when the interrogator makes one of the following statements:
(1) "Pro is female"
(2) "Anti is female"[135]

The Main Game

Turing introduces the Main Game where a digital computer[136] takes the place of Anti and tries to deceive the interrogator and convince her that it is a woman[137]. At the same time, the considerations over masking the physical features and appearances remain unchanged. "The conversations are not to be carried on face to face, of course, since the interlocutor is not to know the visual appearance of either of his two conversational partners. Nor is voice to be used, since the mechanical voice might simply sound different from human voice" (Putnam 1981, p. 33). Again, the best method seems to use an electric typewriter via which the interrogator asks questions and receives the answers. Turing writes:

> We now ask the question, 'What will happen when a machine takes the part of [Anti] in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?' (Turing 1950, p. 434)

Turing claims that this new problem "has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man" (*Ibid.*), since according to the restrictions provided in the Elementary Game, the interrogator is prevented from seeing or touching the other competitors, or hearing their voices or checking their handwriting. "Turing thinks there are other features that distinguish people from machines which might affect the judgment of the interrogator

134 Turing does not indicate if the length of the game is limited or not. However, since he tries to give a practical answer to the question "can machines think?", it seems plausible to believe that the game should finish after a certain amount of time.

135 In Turing's version, Pro and Anti have two labels (Like X and Y). They are located in separate rooms and communicating with interrogator with teletypes. At the end of the game he says either "X is a woman and Y is a man" or "X is a man and Y is a woman". Basically by these statements the interrogator guesses who has been helping her and who has been deceiving her.

136 Turing "argued that limiting machines to digital computers would cause no loss of generality because a digital computer, given enough memory, can mimic any discrete state machine" (Davidson 1990, p. 77)

137 "Turing doesn't mention whether the interrogator is told that a computer has been substituted for the man; and that would surely make a difference to the questioning" (Haugeland 1985, p. 255).

though they should not matter; such features as having a voice, or shining in a beauty contest; the Test removes these features from consideration" (Davidson 1990, p. 79). While the interrogator faces these limits, there is basically no restriction on the questions she can ask, and almost any one of the fields of human endeavor (including questions about everyday life or complicated scientific theories) can be incorporated[138].

Counter-Arguments

Most of the time the counter-arguments against the Turing test conception of intelligence are proposed as thought experiments. Here, I shall address one of the strongest objections against the Turing test conception of intelligence which has been proposed by Ned Block (1981).

Block follows the typical anti-behaviorist argument: He proposes two systems which are alike in their actual and potential behaviors; yet there are differences in the information processing that mediates their stimuli and responses. In order to fit this anti-behaviorist argument into our discussion better, it can be limited to the question of intelligent behavior. According to Block's view, which he calls "Psychologicism", whether or not a behavior should be counted as intelligent "depends on the characters of the internal information processing that produces it" (*Ibid*., p. 5)[139].

He begins his main line of argument by focusing on the Turing test. According to Turing, if a machine passes the Turing test (i.e. the Main Game), it is an intelligent machine[140]. One way to understand Turing's proposal is to think of it as giving the definition of the concepts of intelligence operationally. This operational definition would be something like this: "If a system is given the Turing test, then it is intelligent if and only if it passes the test" (*Ibid*., p.8). If we take passing the test as the definition of intelligence, then it is absurd to ask of a machine that passes the Turing test if it is *really* intelligent: It would be intelligent by definition. Block also, does not take Turing's claim as the definition of intelligence.

---

138 Although, as we saw in the case of Elementary Game, it seems that it takes little reflection and effort to convince one that the best strategy for the one who wants to help the interrogator (Pro) is to give "truthful answers", there is at least one real-life example where an interrogator mistook a human for a computer because the human exhibited what the interrogator thought a superhuman store of knowledge about Shakespeare! (Schieber 1994, p. 70).

139 It should be noted that not every philosopher who argues against psychologicism should be counted as a behaviorist. For example, Putnam (1960, 1967) puts forward a functionalist objection against psychologism.

140 Here "intelligence" is used in a general sense which is something like "the possession of thought or reason" (Block 1981., p. 8).

Block, then, sets up a thought-experiment by which he tries to show that the Turing test is inadequate. Imagine a machine which has the capacity to produce a sensible sequence of verbal responses to verbal stimuli. Block tries to demonstrate that it is possible that the machine is totally lacking in intelligence (given what its internal mechanisms are) despite the fact that it passes the test. In order to grasp Block's argument some terminology should be introduced:

a) Typable: A string of sentences is called "typable" if its members can be typed one after another by a human typist in an hour (that is the length of and actual test in Block's thought experiment).

b) Sensible: A typable string is called sensible if it would be interpreted as a sensible conversation.

The set of typable strings (and therefore the set of sensible strings) is finite. Therefore we can imagine that a machine has a database which contains the set of sensible strings of a particular person. At this stage Block presumes that the machine's programmers try to incorporate "some definite personality with some definite story" in the machine's database in order to get the best results. Let us say the programmers focus on the answers made by a bus driver. They try to load a database with this person's possible answers to different questions. Now the story goes like this. The interrogator types in sentence *A*. The machine looks into its database and picks out those strings that begin with *A*. The machine then randomly chooses one of the members of this subset, and takes out the second sentence from this string, and then types it out as the answer and call it *B*. The interrogator types in sentence *C*. The machine then looks for the strings that start with *A*, followed by *B* and *C*. It chooses one of the possible strings randomly and types out its fourth sentence (which it calls it *D*), and so on[141]. Now if the test is conducted with the real bus driver and this machine, the interrogator will not be able to identify the machine since the participants' answers are (almost) the same.

So Block argues that that this machine has the capacity to emit a sensible sequence of verbal outputs and is thus intelligent according to the Turing test conception of intelligence. However, says Block, it is not. "*All the intelligence that it exhibits is that of its programmer*" (*Ibid*., p.21). This conclusion does not depend on the length of the test (an hour here), and a similar conclusion can be achieved (in principle) for any given period of time. Therefore "the capacity to emit sensible responses is *not* sufficient for intelligence" and the Turing test conception of intelligence, as a sufficient condition, also fails.

141 The machine in this sense is tree-searcher. The first node of the tree is the string indicated by interrogator (A).

What makes Block's argument seem so cogent is the mechanism that the blockhead uses to pass a Turing test, i.e., saving all possible conversations that a particular person (for example a bus driver) may produce in an hour and choose each line of conversation from these saved data. To put it in a nutshell, the machine just copies what a particular person would say in a Turing test. We have a very strong intuition that memorizing and recalling is not sufficient for being thinker. Therefore such a machine is not thinking.

However, I want to argue that with this very mechanism no machine can pass the Turing test. Thus it can be concluded that the machine should be equipped with some other mechanisms in order to pass the test. By adding more mechanisms to what our machine already has, we move away from that convincing original intuition. I mean if our machine uses mechanisms other than memorizing and recalling ones, then we cannot base our conclusion that the machine is not intelligent on the intuition which says a machine that uses only those mechanisms is not intelligent. So we can not easily indicate that the new machine (with a wide variety of embedded mechanisms) is not an intelligent machine. Block's argument is only appropriate for a simple string-searcher machine with a huge database. But that machine, I shall argue, cannot pass the test.

In order to find a way to unmask the machine, the interrogator should come up with questions which consider the machine's ability to impersonate, learn, reason, and perceive, since a string-searcher machine lacks these capacities. The simplest way to question a machine's ability to impersonate a human being is to ask it complex and difficult mathematical questions. Turing is well aware of this problem when he says that a man who is trying to imitate a machine "would be given away at once by slowness and inaccuracy in arithmetic" (Turing 1950, p. 435). For this reason, Block's machine should be equipped with a mechanism which makes it spend different amounts of time on different questions. The human participant spends different amounts of time on different questions and, if the interrogator finds out that one of the participants is taking a fixed amount of time to answer a mathematical question, be it really hard or really easy, she has good evidence to think that the machine is using a memorizing-recalling mechanism. This then helps the interrogator to unmask the machine. Of course our bus driver whose (possible) answers have been loaded to the machine's database may answer "I am not good in math" when he is asked to give the result of 215 to the power 15, and therefore the machine would give the same response without any problem. However, the bus driver certainly can multiply 215 by 15 and this should take more time than subtracting 15 from 215. According to the way Block has described his machine, it takes

the same amount of time for each of these questions to be answered by the machine. Block's machine then would fail to impersonate a human being unless it is equipped with some other mechanisms such as timing.

Questions involving learning can also be incorporated into the test. Block addresses this kind of objection when he talks about new questions but his reply to this objection is too quick and oversimplified. The bus driver may not have any opinion about the latest flood in central Europe, new conflicts in the Middle East, or the result of last night's hockey games in Canada. Block's solution is that the programmers can make their database up-to-date by undertaking the task of "reprogramming periodically to simulate knowledge of current events" (Block 1981, p. 26). However, the interrogator can teach the participants something that they do not know while the test is being conducted and then ask them some new questions concerning these new data. In this scenario Block's machine would be easily unmasked. After all, the Turing test is about verbal communication and should not be limited to Q & As. Imagine that, since our bus driver does not like soccer and knows nothing about the game, the machine would answer "Sorry. I don't have a clue." when it is asked "Which country was the host of the World Cup 2002?" The interrogator then says "Korea and Japan were co-hosts in 2002. Do you know who won the world cup?" and the conversation would be followed like this:

Machine (or the bus driver): Sorry I don't have a clue.

Interrogator: Brazil beat Germany in the final. Can you tell me then who came second in Korea-Japan 2002?

Machine (and not the bus driver): Sorry I don't have a clue.

This is the machine's answer since in Block's scenario the programmers have loaded the Machine's database with all the sensible conversations up to the date of the test. As I depicted, the bus driver does not have a clue about these questions before the test so his corresponding machine would be equally clueless.

It is true that it takes a little effort to program a machine which can answer these questions. Many successful projects have been realized so far but by what we have learned from the field of cognitive AI. However, this cannot be done by simple string-searching mechanisms. The machine should be equipped with an inference-making mechanism to extract some new information from what it has already restored. To make an analogy, a computer with a huge memory can save the result of the sum of many numbers. But due to its limited memory, there always exist numbers whose sum is not in the database. Unless the computer is equipped with a mechanism that can perform the function of addition, there is always a way to check if it can calculate the sum of "new" numbers. Block's machine, then, fails to learn during the test unless it is equipped with some other mechanisms such as learning.

Besides that, human beings associate links with varying strength between certain words. This issue, like the first objection, results in spending different amounts of time on different questions. This fact can be used by a human interrogator in a Turing test (French 1990). Imagine a test in which the interrogator counts the time to get answers:

Interrogator: 'What is a dog?'

Machine or Bus Driver (after *n* seconds): 'It is an animal.'

Interrogator: 'What is a cat?'

Machine or Bus Driver (after *m* seconds): 'It is an animal.'

In the case of human beings, *m* would be significantly less than *n*. But in the case of Block's machine, *n* and *m* are equal because it has no mechanism to determine associative strengths. It shows that "realistic performance required that the computer program have access to an extremely large knowledge base. Constructing the relevant knowledge base was a problem enough, and it was compounded by the problem of how to access just the contextually *relevant* parts of that knowledge base in real time" (Churchland & Churchland 1990, p. 50). Block's machine then fails to answer questions like a human being would unless it is equipped with some other mechanisms such as a mechanism which reconstructs a network of associative links between concepts and a mechanism to retrieve the relevant information.

Big Databases

It may be argued that the abovementioned mechanisms can easily be added to Block's machine without turning it into an intelligent system. It is not a big deal to incorporate mechanisms which help the machine delay in answering some questions, in learning some logical inferences and in reading time and date. The programmers cannot only feed the machine with the bus driver's answers, but they can also measure the amount of time he spends on each answer and add these new data to the database. So when the machine fetches a string it knows how much time it needs to spend before emitting it. And probably the programmers can find similar ways to overcome other difficulties. However we are not dealing with a simple string-searcher machine anymore. Block's argument then loses some of its intuitive strength when its machine is armed with new mechanisms rather than one table-look-up mechanism.

Since what really matters are mechanisms other than memorizing and recalling (such as learning, problem-solving) then "[i]f we ever do make an intelligent machine, presumably we will do it by equipping it with mechanisms for learning, solving problems, etc." (Block 1981, p. 25). It seems that conversations in a typical Turing test can be brought up in a way which question the contestants' learning and problem-solving abilities. Block then should ask the programmers to load

the databases with not only what the bus driver might have answered so far, but with his answers to any possible conversation (for all human beings) in the future. I shall argue that this ultimate-big-data base makes no progress because it is nomologically impossible and navigating in it requires different mechanisms. In other words, expanding the database without adding new mechanisms would result in similar difficulties which help the interrogator probe the machine very fast. This is the subject of the next section.

Combinational Explosion: Logical Possibility Vs Nomological Possibility

As the database gets bigger and better, the access problem gets worse. Exhaustive searches take too much time and heuristics for relevance do poorly. If we restrict our conversation to a vocabulary of 850 words of Basic English and to questions and answers which contain no more than four words and if we set the limit of the number of questions at forty then surely "the total number of all possible permissible games is a large, but finite, number" (Dennett 1986, p. 12). So this confirms that Block was right to think that the number of all "sensible" conversations is finite. However, as Dennett claims, this number exceeds astronomically the number of possible chess games with no more than forty moves. The latter number is around ten to the one hundred twentieth power (10^{120}). So if it takes one second for each conversation to be fed into the database then the amount of time which is needed to load all the possible conversations is massively more than the time from beginning of the universe till now.

To be fair, one should admit that Block is aware of the combinational explosion objection, but thinks that this cannot disprove his argument since all that his argument requires is "that the machine be logically possible, not that it be feasible or even nomologically possible" (Block 1981, p. 30)[142]. Block even briefly addresses the nomological possibility of his machine. He thinks that in contemporary physics, nothing prohibits the possibility of the infinite divisibility of matter. If this is the case in some parts of the universe, then "there need be no upper bound on the amount of information storable in a given finite state" (*Ibid*., p. 32). Block's point is that there are ways to think that his machine is nomologically possible!

In order to show where Block has gone astray, I will make two comments on this argument. First, by focusing on Dennett's position, I will try to complete my claim that Turing's claim was not an attempt to define

142 Searle also insists that speed is strictly irrelevant in his Chinese room thought experiment. A slow thinker should still be a real thinker.

thought in terms of language and I will then conclude with considerations about why nomological possibility is crucially important in this issue. I have already stated why we should not read Turing's test conception of intelligence as an attempt to reduce thought to language.I also will try to argue why it is logically possible that the number of sensible arguments which can be written in an hour is infinite.

Turing Does Not Define Intelligence

If Turing intended to define the concept of 'intelligence' in terms of passing the (unrestricted) test, then Block is right in indicating that passing the test provides both necessary and sufficient conditions of being intelligent. If this reading is correct, then the key move from the question 'can machines think?' to the Main Game, "was to define intelligence operationally, i.e. in terms of the computer's ability, tested over a typewriter link, to sustain a simulation of an intelligent human when subjected to questioning" (Michie 1996, p. 29)[143]. This means that if an agent fails the test it is not an intelligent being because passing the test is the definition of intelligence. However, there are good reasons "for not interpreting the Turing test as an operational definition of thinking" (Moor, 1987).

First, Turing never indicates that he is giving a conceptual or operational definition. Actually at the very beginning of his paper he mentions that such efforts in giving definitions for concepts like 'thought' or 'machine' are fruitless because questions like 'can machines think?' are "too meaningless to deserve discussion" (Turing 1950, p. 433).

Second, the relation between the very first imitation game, what I called the Elementary Game, and the famous Turing test, or what I called the Main Game, a relation which is usually neglected, can shed light on this dispute. In the Elementary Game, apart from lucky breaks, if a man can imitate a woman's peculiar way of thinking then he is an intelligent person. However, failing to do so does not show anything and Turing surely was well aware of that. Many intelligent people may not be able to succeed in imitating the verbal behavior of the opposite sex. This means that "we should allow computers the same opportunity to decline to prove themselves" (Dennett 1986, p. 4). Therefore this is a "one-way test": succeeding in it means having intelligence; failing it means nothing. And that is why for many years philosophical debates over Turing's claim were focusing on "whether or not passing the test would constitute a sufficient condition for intelligence" (French 1995, p. 61).

It is also interesting to see that even the case of logical possibility is not that cogent. I think a wide variety of opportunities can be put forward to show that it is logically possible that the number of typable

143 Searle (1980, p. 423) and French (1990, p. 53) also share this view of operational definition.

strings is not finite. In these cases no computer can memorize all possible conversations due to its limited memory space.
The simplest one, I suppose, is the logical possibility of Zeus typing. Inspired by Boolos and Jeffrey's superhuman creature who can enumerate **N** (the set of all natural numbers) in a finite amount of time (Boolos and Jeffrey 1989), one can imagine that human beings "reactivate" lost typing skills by using some kind of medicine so they can type faster and faster as they move on in the Test. It is logically possible, then, to imagine that one day (perhaps long before the realization of Block's machine), a human interrogator can proceed with an infinite number of sensible conversations (in an hour). Then Block is pushed to argue (as he did for other reasons) that the machine's memory can be expanded unlimitedly. This again shows that Block's machine in its original form cannot pass the test. And now that its memory is not finite we need to introduce a new mechanism to navigate in this infinite data base. These points show that as the machine improves we get distanced from the original intuition that it is not a thinker.

Summary and Conclusion

In this paper, I introduced the TTCI. I stated that Turing describes his test through two stages, what I called the Elementary Game and the Main Game. I mentioned that Turing is not clear on the relation between the two games. But as some commentators (e.g. Dennett) have argued, analyzing the Elementary Game and its relation with the Main game can shed light on TTCI. I also stated that there are two central questions in Turing's paper one thought and one concerning language (namely 'Can machines think? and 'How good would a machine perform in an imitation game?'). Turing claims that the first one can be replaced by the second (but he does not bring any argument for the validity of this substitution).

I depicted Block's Blockhead as a counter-example against the TTCI. I argued that, in contrast to what Block has said, this machine (in its original format) cannot pass the Turing test and therefore is not a treat to it. In addition to this especial response, I have a general reply to all of counter-arguments against the TTCI which are based on a thought-experience. In this section, I shall explain this general reply. This reply is crucial to understanding my interpretation of the Turing test.

Quine (1951) has persuasively argued that that there is no (sharp) distinction between analytic statements whose truths are grounded in meaning independent of matters of fact and synthetic statements whose truths are grounded in fact. The immediate conclusion of this argument is that "no statement is immune to revision" (*Ibid*. p. 43). Applying this conclusion to our discussion means that any attempt to define intelligence is open to logically possible counter-examples. Once Quine's

argument has been accepted, one does not need to read weird scenarios in which a non-intelligent agent passes the Turing test in order to accept that there are cases in which Turing's claim does not work.

Even a monkey can pass the test out of pure chance. Imagine a lucky monkey who is sitting in one room competing with an intelligent human being in a typical Turing test. It plays with the key board and happens to produce the right answer at the right time. This counter-example, as it seems to me, is even more possible than the aforementioned thought experiences: it is clearly both logically and nomologically possible. But if such a thing happens, we are not inclined to say that Turing's claim has been falsified. Even the position of Turing's claim in our web of beliefs will not change. Turing has claimed that passing the test gives us justified reasons to believe that the agent can think. This claim, like any other claim about thought and intelligence, is not immune to revision. The point that should be considered is how nomologically possible the proposed counter-examples are. Surely, Turing was well aware of the possibility of non-intelligent objects/creatures which pass the test by sheer luck. Passing the test is evidence and the participant's answers show an apparent knowledge of the world. "It would be too much to suppose all this is an accident" (Davidson 1990, p. 82); while the possibility of accident always remains.

Counter-examples in which the number of memory units we need are more than the total number of the atoms of the universe and the time we need is more than the seconds that have passed since the Big Bang are not really threatening counter-examples. Any other claim about intelligence would be open to similar nomologically impossible counter-examples. Such claims cannot change the position of a particular belief in our web of beliefs. It is logically possible that each time we examine the chemical composition of water a devil spirit deceives us. This is not a very good argument in order to conclude that water is not made up of hydrogen and oxygen. We have relatively strong justified reasons to believe that water is made up of hydrogen and oxygen.

The monkey example shows that there is a very easy way to acquaint the reader with a situation in which an agent, lacking intelligence, passes the test. However this example says nothing more than what Quine has already claimed. "Since Quine's (1951) attack on the analytic/synthetic distinction, it is commonly held by philosophers that the attempt to provide necessary and sufficient conditions on almost any concepts- i.e., to analyze or define almost any concept- is doomed to failure" (Jacob 1997, p. 21).

In order to criticize the TTCI one must bring up issues which have been neglected by Turing. Having a proper history and having the right kind of causal interactions with the environment could be among these issues.

References:

Block, N. (ed.) 1980. *Reading in Philosophy of Psychology*, Vol. 1. Cambridge: Harvard University Press.

Block, N. 1981. "Psychologism and Behaviorism", *The Philosophical Review* **90**. 5-43.

Boolos, G. S. and Jeffrey, R. C. 1989. *Computability and Logic*. Cambridge: Cambridge University Press.

Churchland, P. M. & Churchland, P. S. 1990. "Could a Machines Think?" in Churchland & Churchland (1998). 47-63.

Churchland, P. M. & Churchland, P. S. 1998. *On the Contrary*. Cambridge: The MIT Press.

Davidson, D. 1990. "Turing's Test" in Davidson (2003). 77-86.

Davidson, D. 2003. *Problems of Rationality*. Oxford: Oxford University Press.

Dennett, D. C. 1978. *Brain Storms: Philosophical Essays on Mind and Psychology*.

Dennett, D. C. 1986. "Can Machines Think?" in Dennett (1998). 3-30.

Dennett, D. C. 1998. *The Brainchildren*. Cambridge: The MIT Press.

DeRose, K and Warfield, T. A. (ed.) 1999. *Skepticism: A Contemporary Reader*. Oxford: Oxford University Press

French, R. M. 1990. "Subcognition and the Limits of the Turing test", *Mind* **99**. 53-65.

French, R. M. 1995. "Refocusing the Debate on the Turing Test: A Reply to Jacquette", *Behavior and Philosophy* **23**. 61-62.

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge: The MIT Press.

Jacob, P. 1997. *What Minds Can Do*. Cambridge: Cambridge University Press.

Michie, D. 1996. "Turing's Test and Conscious Thought" in Millican and Clark (1996). 27-51.

Millican, P. and Clark, A. (ed.) 1996. *Machines and Thought*. Oxford: Clarendon Press.

Moor, J. H. 1987. "Turing Test" in Shapiro (1987). 1126-1130.

Putnam, H. 1960. "Minds and Machines" in Putnam (1975). 362-385.

Putnam, H. 1967. "The Nature of Mental States" in Block (1980). 223-231.

Putnam, H. 1975. *Mind, Language and Reality. Philosophical Papers,* Vol. 3. Cambridge: Cambridge University press.

Putnam, H. 1981. "Brains in a vat" in DeRose and Warfield (1999). 27-42.

Quine, W. V. 1951. "Two Dogmas of Empiricism" in Quine (1961). 20-46.

Quine, W. V. 1961. *From a Logical Point of View*. Cambridge: Harvard University Press.

Searle, J. R. 1980."Minds, Brains and Programs", *Behavioral and Brain Sciences* **3**. 417-424.

Shapiro, S. C. (ed.) 1987. *Encyclopedia of Artificial Intelligence*. New York: John Wiley and Sons.

Shieber, S. M. 1994. "Lessons from a Restricted Turing Test." *Communication of the ACM* **37**.

Turing, A. M. 1950. "Computing Machinery and Intelligence", *Mind* **59**. 433-460.

ROBOTS, ZOMBIES AND FREEDOM:

AN ESSAY TRYING TO END THE DISCRIMINATORY USE OF THE ABHORRENT TERM 'ZOMBIE' FOR OUR FULLY AUTONOMOUS SERIOUSLY QUALIA IMPAIRED FELLOW HUMANS

TILLMANN VIERKANT
Philosophy Department David Hume Tower EH8 9JX Edinburgh

Email: T.vierkant@ed.ac.uk

Abstract:

The paper investigates the connection between freedom and consciousness. It argues that zombies are for questions of practical philosophy indistinguishable from conscious humans. This has the implication that phenomenal consciousness can only be relevant for practical philosophy, if the hard problem is not as hard as its proponents want it to be. The paper then explores in which way consciousness might be relevant for freedom and in which way phenomenality might be connected to relevant forms of consciousness.

Acknowledgements:

For helpful comments I want to thank Joelle Proust and Susan Hurley. This paper would not have been possible without the generous support of the Max Planck Institute for Human Cognitive and Brain sciences and the Volkswagenstiftung.

Zombies on Trial

Imagine the following scenario: Judge Anthony Just has more than enough evidence that the vicious murderer in his court is mentally sane and knew exactly what he was doing when he killed his victims. He likes to see himself as a rational, superior being who can take the lives of normal people simply for his own pleasure. He has no scruples about perceiving himself in that way, and there seems to be no pathological reason for his emotional coldness. He displays normal emotional behaviour towards his family, which he seems to hold in high esteem because they contribute to what he would call his good life, but this does not prevent him from mercilessly taking life where ever it seems useful to him to do so. Our judge is therefore only moderately interested when the defence calls a new witness. The philosopher and possible-world expert Prof. Calmos is led in. The defence tells the judge that Prof. Calmos is the inventor of a very interesting tool called the consciousness detector, and that he would like to use this detector to carry out a very important test on the accused. Judge Just allows the test. Prof. Calmos takes his tool (which looks very much like a conventional hair dryer) and holds it to the head of the accused. After a short investigation he exclaims: 'There is no evidence of phenomenal consciousness here; this person is a philosophical zombie'.

The defence jumps up triumphantly and cries: 'The defendant is innocent! He was not conscious of what he was doing, because he has no phenomenal consciousness at all. Therefore, he cannot possibly have had conscious control over his actions and thus, he cannot be held responsible for them. Once more, I demand freedom for my innocent defendant!'

Judge Just is very surprised by this turn of events, so he adjourns the court and retires quickly to his room in order to think through what relevance the discovery may have. This paper wants to retire with him and pose the same question. The paper is convinced that Judge Just actually is having a bad dream, because the scenario he finds himself in is on the view of this paper probably impossible. This is because the author of the paper does not believe that such a strange creature as a philosophical zombie could exist. But even if it should be the case that zombies could be made conceptually possible, this comes at a high price.

In order to make the idea of zombiehood coherent, one has to narrow the scope of the term consciousness so dramatically that it becomes almost unrecognisable. As soon as we are trying to make it a little richer, the concept becomes contradictory. Judge Just's nightmare translated into philosophical language becomes a neat reductio ad absurdum of some strange ideas about phenomenal consciousness. If the argument succeeds, this does not show that there could be nothing like the hard problem of consciousness (Chalmers, 1996)[144], but it does show that this problem (if it should indeed be as hard as it's proponents think it is) does not need to be solved in order to understand the human capabilities to act intentionally, freely and responsibly. Surely, if we can explain these capabilities in a naturalistic framework and might even be able to artificially build agents who have them, it would not be tragic, if we could never explain the blueness of blue.

Zombies and Free Will

Zombies are weird creatures. The traditional philosophical zombie lives in a world that is just like ours, apart from the fact that there is no phenomenal consciousness. They are supposed to be a philosophical argument for the non-reducibility of the phenomenal to the functional. This argument runs roughly as follows: As zombies are conceivable, and because they are exactly like us apart from their lack of phenomenality, phenomenality cannot be identical to anything that zombies have - which is to say that phenomenality in us is not describable purely in terms of anything else in us. The possibility of zombies would be a fatal blow for naturalistic theories of consciousness, because if it were true that zombies could exist then it would be true that the explanation of consciousness poses an unsurmountable hard problem, because consciousness could not be identical to anything in a materialistic world. After all zombies are identical with respect to all material properties of humans and still do not have phenomenal consciousness. There is a lot of debate about the conceivability of zombies and it seems to me that there have been quite a few convincing arguments that such creatures could not possibly exist (Balog, 1999; Dennett, 1995), but the debate persists and one aim of this paper is to open up another front in the zombie wars, by showing that, if zombies could exist, then this would only show that the problem of the explanation phenomenal consciousness

144 This is e.g. Chalmers' claim that the empirical sciences can solve many problems associated with consciousness solely with traditional empirical methods, but will find it hard (read impossible) to solve the fundamental riddle of how the grey matter of our brains can produce the phenomenal richness of our mental lives. It seems impossible to imagine, say, that our phenomenal impression of the blueness of the sea is just a specific function of a neural network in our brains.

might be 'hard' but at the same time less important for an understanding of human mental life than it may seem. The main aim, though, is a clarification of terms for practical philosophy. In the debate on free will and agency that is going on at the moment, the role of phenomenal consciousness for action control and autonomy is hotly contested. Here, I want to clarify what we can sensibly mean in this debate by phenomenal consciousness and what we cannot mean. I want to show that the hard problem of consciousness is not one for practical philosophy.
In the remainder of this section I will explain in a bit more detail what I mean by saying that zombies would be autonomous agents. Afterwards I will explore why that clashes with some assumptions of practical philosophy in order to finally get a better grip on the relevance of consciousness for practical philosophy.
That zombies are free and autonomous agents seems to follow from a very simple argument: If zombies are a possibility, then this shows that an unconscious being can have just as much control over its behaviour as a conscious one. It shows as well, that there are sufficient functional mechanisms in our world to explain the control of our behaviour without referring to phenomenal consciousness. After all, we have everything that zombies have and that is, as zombies show, sufficient for the control of behaviour. Phenomenal consciousness, on the other hand seems to provide the functional level with an experiential feel (a 'what is it like' dimension), without changing anything in the behavioural world.[145] As free will surely is all about the right kind of control over our actions[146], zombies are just as free as we are. As it seems quite plausible to claim that freedom of the will is a sufficient condition for the ability to be responsible for one's behaviour, zombies seem to be indistinguishable from ourselves from the point of view of practical philosophy. They are nothing like automata or robots, which can perform amazing cognitive operations but are ultimately non autonomous agents to which the language of morality could not be applied, but they are people like you and me with a pretty strange form of colourblindness.
One might object here that zombies might have the same behavioural control, but that this control is not the right kind of control, because it is not in the right way that they control their behaviour. What is lacking is the mental subject that is the controller. Instead in a zombie the control is performed by sheer mechanisms. These, so one could argue, do not amount to a person that could control in the relevant sense.
But this argument does not work: To see why, it is helpful, to look at the position of one of the staunchest defenders of zombiehood David

---

145 At best there could be some form of double causation going on.

146 This holds true for the compatibilist control through reasons as well as for the behavioural control flowing from the capability to do otherwise cherished by the libertarians.

Chalmers, who rightly criticises Goldman for arguing that zombies could not have beliefs. As Chalmers correctly writes:

> 'Zombie Dave's beliefs may not be colored by the usual phenomenological tinges, but it seems reasonable to say that they are nevertheless beliefs. Beliefs, unlike qualia, seem to be characterized primarily by the role that they play in the mind's causal economy. (To illustrate the difference, note that it seems *coherent* to be an epiphenomenalist about qualia, whether or not one finds the position plausible; but there seems to be something *conceptually* wrong with the idea that beliefs could be epiphenomenal.) So qualia-free believers like Zombie Dave are quite conceptually coherent, and qualia don't seem to be an essential part of our concept of belief"…Chalmers ends his comment with the sentence … 'there may not be any tenable middle ground between functionalism and epiphenomenalism." Chalmers commentary in (Goldman, 1993).

I could not agree more with this conclusion, but it seems to me it has not been emphasised enough what that means for the significance of the hard problem and for the role of phenomenal consciousness in practical philosophy.

What Chalmers says for beliefs counts as well for all the other mental items that might be crucial in determining the freedom of an action. It seems conceptually wrong to imagine that they could be epiphenomenal. If they were, they would not make us free! But this means that the objection raised about the form of the relevant control cannot be valid. What makes us free is something that exists in Zombies as well, so if they control their behaviour in the wrong kind of way so do we. This is to say that my claim is supposed to be valid not only for compatibilists but also for all the libertarians who believe that the mental has a specific causal capacity and that it is this specific capacity that allows us to be autonomous.

In my view this means that if zombies were to exist, our judge should have no qualms about sentencing his accused zombie, because he was just as free as a conscious copy of him would have been. If the argument is correct then phenomenal consciousness may still be interesting in its own right, but it is not necessary for an understanding of our capacity to act intentionally or freely. Whoever claims that it is and believes in the possibility of zombies seems to commit himself to a statement that is ultimately incoherent.[147]

147 John Searle and Alvin Goldman (Goldman, 1993; Searle, 1993) would probably be good candidates for such incoherencies.

So far, it seems there is no problem. Zombies are just like us in all relevant aspects, so that they are fully responsible beings, capable of understanding the consequences of their deeds. The ‘zombie on trial’ thought experiment seems to make it perfectly clear that for all the questions of practical philosophy phenomenal consciousness is strictly irrelevant. But there is a problem and here it is: David Velleman (Velleman, 1992) has a wonderful thought experiment that clearly shows that we do not feel that phenomenal consciousness is irrelevant for intentional actions at all. Imagine the following: you are disenchanted with an old friend of yours who recently started to make comments that seem cynical and abhorrent to you. You had decided to end the friendship for quite a while when you meet him again[148]. On this particular occasion ending the friendship is not on your mind, you are just glad to see him again. But while you talk to each other you get more and more heated in the discussion and finally insult each other so badly that it leads to the breaking of the friendship. Afterwards you realise in a calm moment that your prior thoughts about your friend had given your arguments such a nasty and unfair edge without you ever being aware of it that your friend had little choice but to end the friendship. You feel like you missed a great chance to mend fences with your friend and are very sad that your remarks led to the final breaking of the friendship. I will refer to this scenario as ‘friendship’ from now on.

On a standard philosophical account of action, that leaves out phenomenal consciousness, your action, quite to the contrary to your own evaluation of it, was a prototype of a free action. According to such an account, an action is intentional, if your intention is causally influential in bringing about the action. In this case, you certainly had formed the intention to end the friendship and this intention was certainly influential in bringing about the intended result, even though you were not aware of it at the time. It is not a case of the infamous deviant causal chains either. Deviant causal chains are about special cases of behaviour that the standard story admits to, but holds that they are very rare. In such cases an intention is causally influential in bringing about an action, but not in the way it is supposed to. Donald Davidson’s (Davidson, 1973) example of a climber who gets so nervous, because he knows that he would like to

148 Velleman’s original example does make slightly stronger assumptions here. He believes that the thought example would go through, even if you had never been aware of your intention to sever the friendship. I have opted for the slightly weaker assumptions, because I think they are enough to prove the most important point that the agent was not involved in the action sufficiently for it to count fully as her action. My way of drawing up the story also highlights the importance of online control which I discuss later in the paper.

let go of the rope, that he does in fact let go is the standard example. The 'friendship' scenario is importantly different though. In this scenario it is not the sheer existence of an intention that triggers in a very unusual roundabout way involuntary behaviour, but the intention influences the action very much in the normal way i.e. in the way that intentions control our behaviour in normal cases of actions. But if this should be true, then on the standard story the described example should be a prototypical intentional action.

But certainly this cannot be the correct interpretation of the thought experiment. It is true that you had intended to end the friendship and it is true as well that this made you nastier than you would have been otherwise, but you did not *want* to be nasty to end the friendship! This is why you describe the situation passively. It was your intention and not you that severed the friendship. In the conversation you did not want to do anything about the friendship, that it happened nevertheless was certainly a consequence of your prior intentions, but not one that you were conscious of. You cannot claim that you had nothing to do with the action, but it did not have your full approval either. It was an accident that it happened. Surely such a situation differs massively from one where you consciously make the remarks in order to provoke a breaking of the friendship. You can't escape responsibility in the first case, because your action was after all caused by some intentions that you had consciously entertained before, but you are obviously not responsible for such an action in the same way as in the second case, where you are fully aware of what you are doing.

But if this is what we have to conclude from 'friendship', then zombies are never responsible in the second sense, because they are never aware of their actions. They control them, they have beliefs about them, but they have no awareness of their actions. So looking at it from this angle it seems Judge Just should let our zombie off. His behaviour was certainly controlled by his unconscious intentions, but he wasn't aware of having them. Worse, we cannot even make him responsible in the same sense as we can make the actor in 'friendship' responsible, as we make this person responsible because he did consciously *form* an intention to end the friendship. The zombie, on the other hand, did no such thing. The intention was just there, formed unconsciously and without him ever being aware of it.

This leaves us with an obvious inconsistency. Zombies seem to have what is sufficient for responsibility i.e. the right kind of control over their behaviour, and at the same time it seems as if they cannot be responsible, because they are not aware of what they are doing. To escape this dilemma we could do one of three things. One could do what I will ultimately recommend and give up the idea of a consciousness that allows for zombies as useful for explaining anything about practical issues. But one could as well deny that the Velleman example shows that

not being aware of an action rules out autonomy. Finally, one could opt for claiming that control is not what matters for responsibility. In the following section I will discuss the second option. The third option, it seems to me, is so implausible that I am quite willing to let it stand for the moment. Going down that road implies giving up one of our deepest convictions about the structure of responsibility: the conviction that we can only be responsible for an action if we could have done otherwise, if we had wanted to do otherwise.[149]

Is Vellemann's Example about Consciousness at all ?

In favour of the second option is certainly the fact that 'friendship' was not designed to show anything about conscious awareness. In fact, Velleman even argues explicitly against Carl Ginet's account of action, e.g. (Ginet, 1990) who defines simple mental actions[150] as actions that have a 'phenomenal actish quality' i.e. feel like an action to the actor. According to Velleman, this account fails to capture what we mean by our concept of action, even if it should rightly describe what happens in all the events that we call actions. This is because Velleman believes that on Ginet's account causal control of actions by the actor is ultimately illusory. The modern standard conception of action on the other hand holds that actions are actions because they are caused by the intentions of the agent. According to Velleman, this theory has a considerable advantage over Ginet's, because it captures one import aspect of our folk concepts of agency i.e. that we are convinced that we cause our actions.
Nevertheless, Velleman wants to show that the standard story of intentional action is missing something that is crucial for our understanding of human autonomy as well. This something that the standard story is missing is something that according to Velleman is understood in the recent renaissance of philosophical theories of agent causation. The standard story does not seem to have a good answer to the question why a desire belief calculus, which according to the standard story constitutes the agent, is not really only a passive mechanistic device which does nothing to merit the term agency. Agent causation on the other hand may be extremely implausible, but it takes seriously the

149 This condition is obviously not the standard condition in which incompatibilists believe. Here it is the weak claim that we can only be responsible for our actions, if what we want is influential in what we do. It says nothing about being able to do otherwise under identical conditions. There are nevertheless people who have defended such a proposal. For example see (Owens, 2000).

150 which are at the same time the causal beginnings of all complex actions.

strong intuition that we are more than our desires and beliefs and that we can use this extra weight of our agency to influence our decisions. For Velleman his example is not about awareness, but about the fact that the agent was not involved in making the decision to be nasty. Velleman then gives an account of what this mysterious agent could be. What he proposes has very little to do with an incompatibilist agent causation account. For Velleman, the missing agent is nothing else but a specific cognitive function, the desire to act rationally. Velleman borrows here from the Frankfurtian (Frankfurt, 1971) solution of the freedom of the will problem. Free Agency is constituted by a specific desire, but contrary to Frankfurt, for Velleman it is not an identification by the agent with a higher order desire that constitutes agency, but a specific desire i.e. the desire to act rationally is functionally identical to the agent.[151]

This is a strong account of essential features of what it means to act freely, but in giving this account, Velleman overlooks the fact that he does not solve the problem of his own thought experiment. It seems quite possible that the actor in 'friendship' would feel that he did not actively and freely end the friendship, even if he had decided beforehand that it would be rational to end the friendship, thereby deciding an internal motivation conflict between nostalgia about the good times and annoyance about the remarks the friend makes every time they meet. He might even be relieved that it happened that way, because he might still feel that it is better all things considered, but might well insist nevertheless, that he didn't intentionally do what he did, because at the time he did not want to act for these reasons.

This turns the argument that Velleman used against Ginet on its head, it seems now that Velleman probably describes an existing cognitive function i.e. the desire to act rationally, but that this desire is not sufficient to understand our *concepts* of agency, because it lacks the dimension that we only feel that we do something intentionally if we are at the time of action aware of what we are doing. This dimension is lacking in Velleman's desire account, but it is captured in Ginet's 'actish phenomenal quality', which guarantees that we are in control of what we are doing at the time we are doing it.[152] Ginet's account guarantees that we, as it were, are not only setting things up, but are controlling them online in real time via our phenomenal consciousness. Velleman might not have wanted to make a case for the necessity of phenomenal consciousness for agency, but it seems that he succeeded in doing so

151 Velleman is concerned with action in general, while Frankfurt is concerned with freedom of the will. So Velleman uses the involvement of the agent on a more fundamental level than Frankfurt. To reintroduce the distinction between intentional and free actions, Velleman would need an extra distinction between e.g. a merely strategic involvement of the agent in an action and actions that are motivated by the agent.

152 This obviously is no proof that Ginet's theory is right (in fact I am convince it is wrong)! It just shows that Ginet captures something with his theory that an appropriate conception of action will need to account for.

nevertheless. This brings us back to our zombies. It seems we still don't understand how a zombie could be responsible, even though we know that they must be!

The Missing Link: Self-Understanding and Consciousness

The last section closed on the note that a desire to act rationally could well be imagined to work unconsciously and that we can therefore not use it to solve the ‘friendship’ scenario. But perhaps we were slightly to quick in conceding that possibility. Since John Locke (Locke, 1995) there has been a long and powerful tradition within philosophy that equates consciousness with some form of self awareness. Now, if this is the conception of consciousness that we are going to work with, then the self function Velleman describes might entail consciousness. The presence of the much vaunted desire to act rationally in a cognitive operation which is according to Velleman identical to the functional agent might as well entail that this process becomes conscious.

This is one explanation why we struggled with the idea of a free zombie, because it means that it is impossible to conceive of zombies, because they have contradictory attributes. As they need to be identical to conscious beings in terms of their functions, it is bad news for the zombies if one cognitive function entails consciousness.

One strategy to avoid this conclusion for the zombie defendant would be to claim that this idea about consciousness cannot be sound, because the behaviour of the person in ‘friendship’ was in line with what he wanted all things considered, but it still did not feel phenomenally as if it had been an action. This means that the desire to act rationally cannot be what we are looking for when looking for conscious behaviour. This is an excellent observation but a bad comeback for the zombies, because the observation only shows that to set up a specific behavioural tendency is not the same as controlling that behaviour while it is happening. What is lacking in the example is the *online* control by the agent that Ginet could account for. But obviously this kind of online control does not require a strong notion of phenomenal consciousness but is quite plausible in a model that only operates with access consciousness. As soon as we add the online condition to the equation between agent controlled behaviour and conscious behaviour, it becomes intuitively quite plausible and resistant against criticisms using examples like the Velleman case.[153]

153 Only by adding this condition can we begin to understand as well the challenge of recent findings in the cognitive sciences that seem to show that there is a lot less conscious online control than we tend to think in our folk psychology. The implications of this are only slowly emerging. This process is obstructed to a considerable degree by

Is it the Right Consciousness?

The best and obvious counter argument for the zombie defendant is that the understanding of consciousness in the previous section is by no means undisputed. There is almost nobody who would deny that some form of self awareness or subjectivity is central for an understanding of some form of consciousness, but it does not seem clear what that entails. There is, for example, Ned Block's famous distinction between access consciousness and phenomenal consciousness. A state is access conscious, according to Block, if in virtue of having this state, a representation of its content is inferentially promiscuous and poised for rational control of action and speech. Having access consciousness does not entail that one has phenomenal consciousness, there might be no 'what is it like' dimension to a state that is access conscious.[154] Now, one could hold, that in the Velleman case there is access consciousness but no phenomenal consciousness. This would show that access consciousness is not enough for us to rate a behaviour as intentional. But this is not a very promising strategy because, as shown in the last section, it is quite plausible to hold that access consciousness involves online control by the agent so that there is no access consciousness in the Velleman example. In fact, the very point of the example is that there is no global availability of the influence the intention is having on the behaviour of the actor. The definition of access consciousness is actually not that dissimilar to the functional actor that Velleman describes himself.
More promising seems the strategy that admits that the Velleman example is not suitable but holds that it is quite possible to imagine a case where there is access consciousness, but where the actish phenomenal quality is absent and that we would still describe such a situation as one where agency was lacking. This objection is build on the critic that Velleman overlooks the importance of control at the time of action, but it goes further in arguing that not any type of control will do.
In order to understand what such a situation could look like we have to introduce another very strange philosophical creature, the super-blindseer. Blindsight is a fascinating phenomenon in neurophysiology - see e.g. (Weiskrantz, 1997). Blindsighted patients are cortically blind, but

misunderstandings about the nature of the challenge. Many people still believe that these experiments are about the philosophical debate between libertarians and hard determinists or compatibilists. This confusion arises because the role of the phenomenal in action control is never explicitly made clear.

154 This is the definition that Block gives in his entry on consciousness in (Guttenplan, 1994).

are still able to guess above chance level for example whether something is moving in their blind field, when prompted to do so. Nevertheless, they firmly deny having any phenomenal experience of the moving object. Super-blindsight is a philosophical radicalisation of the empirical phenomenon. There is no such thing as super-blindsight in the real world and, as in the case of zombies it might well be impossible.[155] It is a merely philosophical tool that is used to illustrate the difference between phenomenal and functional consciousness. A super-blindseer is importantly different to the empirical phenomenon in that he is supposed to be able to prompt himself to guess and he is supposed to guess not only about chance level, but at hundred percent. The super-blindseer in this way has full access to the fact that there is a moving object without the experiential 'what is it like' dimension.

Now imagine Fred:

Fred is a super-blindseer. He is cortically blind in his left visual field, but he is able to 'guess' when something is moving in this field. Fred is controlling a button which is connected to an explosive situated in his blind field. 'Guessing' that the doctor he does not particularly like is entering the room, Fred detonates the explosive, although he does not have any phenomenal experience of the reason for executing the movement. Nevertheless, if asked afterwards why he detonated the explosive he says: well I knew he was coming, I don't like him and that is why I pushed the button.

It does not seem to be sensible to argue that Fred was not autonomous in his decision to detonate, because he was not aware of the doctor. For the question of freedom it seems completely irrelevant how Fred knew that the doctor was coming, what counts is that he did know and that he was at the time of the action's execution knowingly in full agreement with what he was doing. For Fred, his action had not the actish phenomenal quality of acting for a reason, but he knew that he was controlling his behaviour. He knew as well that he could have stopped it, if he wanted to stop it, but freely chose not to.

But if Fred is free, this means that access consciousness is all that matters for freedom. We have finally come full circle and can see why the combination of 'friendship' and the conceptual possibility of zombies is not as paradoxical as it first seemed. Freedom does require consciousness, but it does not require phenomenality as understood by the advocates of the division between access and phenomenal consciousness. We do not need to solve the hard problem in order to understand what makes us free and responsible beings. Zombies and

---

155 They were introduced by Block, because they seemed to him conceptually less problematic than zombies. See e.g. his entry on consciousness in (Guttenplan, 1994). Others were not so convinced that superblindseers are in this respect superior to zombies, e.g. (Dennett, 1995)

super-blindseers, if they are a possibility, would be fully autonomous human beings.

How could phenomenality be important for freedom?

Up to now, it was argued that phenomenal consciousness is irrelevant for all practical purposes. It was stressed nevertheless that this is only true on a certain understanding of phenomenal consciousness. Obviously, there are other understandings which do not allow for zombies which are not ruled out by the argument. These understandings can be divided into two main categories.

On the one hand, there are understandings of phenomenal consciousness as a special form of cognitive processing tied to certain discriminatory abilities distinct from access consciousness (e.g. (Lamme, 2003) and on the other hand, there are understandings of phenomenal consciousness as essentially identical with some form of access consciousness.

If access consciousness entails phenomenality then it would obviously be the case that on my account, phenomenality is a necessary condition for autonomy. If phenomenal consciousness amounts to some form of discriminatory ability distinct from access consciousness, then lacking this ability can only be of very limited importance for autonomy. It would mean that a being without this ability could only be responsible for behaviour that does not require the ability. Such an understanding literally reduces lack of phenomenality to some kind of colour-blindness. If this is the right account of what phenomenality amounts to then it is only in the same sense important for freedom as colour-blindness is.

The case is similar if phenomenality is understood as the epistemically special situation of the first person. If a being were able to know that a particular action is forbidden in a third person description, but would not be able to recognise that she is about to perform that action, then she could not control performing that action and would therefore not be responsible for it. She would be, as it were, blind to knowledge by acquaintance.

Never Mind the Hard Problem

Philosophers always emphasise that philosophical zombies have very little to do with the ‘night of the living dead”, but they seem to assume that there is one important similarity between the two kind of zombies

nevertheless. From the fact that philosophical zombies have no phenomenal consciousness it is normally explicitly or implicitly inferred that they are, like the Hollywood creatures, only automatons, heteronomously controlled without any will of their own. I have shown here that philosophical zombies would be free human beings, who would be in every practically relevant aspect identical to us. This in turn means that phenomenal consciousness understood in such a way as to allow for the possibility of zombies is for all practical purposes irrelevant![156]

There are many understandings of phenomenal consciousness that do not warrant such a strong conclusion. On some understandings phenomenality, is just one tool for the control of our behaviour. On other understandings, it is inseparably linked to our capacity to act freely, but all these understandings have one thing in common: They do not allow for the possibility of zombies. For all these understandings the hard problem is not so hard after all and is therefore tractable using the methods of the empirical sciences.

I have not shown that consciousness is not important for freedom. Quite the contrary: what I have tried to do here is to show that an understanding of consciousness is central for our understanding of freedom, but not as a mysterious hard problem, but as a specific control function. One implication of this is that it should be no longer possible to claim that automatic robotic agents (the poor man's zombies) could not be autonomous and responsible agents on principle for the simple reason that they could never have phenomenal consciousness.

Literature:

Balog, K. 1999. Conceivability, possibility, and the mind-body problem. Philosophical Review, 108, 497-528.

Chalmers, D. 1996: The Conscious Mind. Oxford: Oxford University Press.

Davidson, D. 1973. Freedom to act. In T. Hondrich (ed.), Essays on freedom of action. London: Routledge.

Dennett, D. 1995. The unimagined preposterousness of zombies. Journal of Consciousness Studies, 2, 322-326.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. Journal of Philosophy, 67(1), 5-20.

156 At least qua its phenomenality. It might still be the case that in our world all responsible beings are phenomenally conscious, but they are responsible qua some functional property and not qua their being phenomenally conscious.

Ginet, C. 1990: On Action. Cambridge: Cambridge University Press.

Goldman, A. 1993. The psychology of folk psychology. Behavioural and Brain Sciences, 16, 15-28.

Guttenplan, S. (ed.). 1994. A Companion to the Philosophy of Mind. Oxford: Blackwell.

Jackson, F. 1986. What Mary didn't know. The Journal of Philosophy, LXXXIII(5), 291-295.

Lamme, V. 2003. Why visual attention and awareness are different. Trends in Cognitive Science, 7(1), 12-18.

Locke, J. 1995: An Essay Concerning Human Understanding.: ILT Digital Classics.

Owens, D. 2000: Reason without Freedom. London: Routledge.

Searle, J. R. 1993. The problem of consciousness. In C. F. Symposium (ed.), The Experimental and Theoretical studies of Consciousness. Chichester: John Wiley.

Velleman, D. 1992. What Happens when someone acts? Mind, 101.

Weiskrantz, L. 1997: Consciousness Lost and Found. Oxford: Oxford University Press.

LANGUAGE AS A COGNITIVE TOOL

COMPUTATIONAL MODELS OF A PHILOSOPHICAL TOPIC

Marco Mirolli, Domenico Parisi

Institute of Cognitive Sciences and Technologies, CNR, 44 Via San Martino della Battaglia, 00185 Rome, Italy (marco.mirolli@istc.cnr.it; domenico.parisi@istc.cnr.it)

Language and cognition: the 'received' view and its critics

What is the role of language in human cognition? This is one of the most important questions we have to address if we want to understand the human mind. The standard view of classical cognitive science can be summarized with two statements: (a) cognition is, generally speaking, 'linguistic' in itself, in that it is the manipulation of language-like structures (propositions) according to formal rules; (b) the function of natural language is just to express these language-like structures; therefore, natural language does not affect cognition in any substantial way.

The view of cognition as symbol manipulation is at the very heart of classical cognitive science, constituting the common assumption of at least three of the sub-disciplines that gave birth to cognitive science: artificial intelligence (the symbol system hypothesis; Newell and Simon, 1976), cognitive psychology (the language of thought hypothesis; Fodor, 1975), and cognitive-science-related philosophy of mind (i.e. computationalism; Putnam, 1963).

If one considers cognition as fundamentally linguistic, then there is no reason for viewing language as anything more than a very complex and powerful communication system. And, in fact, this view of language has been seldom if ever questioned inside traditional cognitive science.

The basic assumptions of classical cognitive science, however, have been questioned over the years from several perspectives. For example, a number of philosophical arguments have been put forward against the view of cognition as symbol manipulation (see, for example, Dreyfus, 1972; Dennett, 1978; Searle, 1980; Churchland, 1981). But in the absence of concrete alternative proposals advocates of the view of cognition as symbol manipulation could still claim that their hypothesis was "the only game in town" (Fodor, 1975).

In the last twenty years a number of such alternatives have been proposed. The first one was connectionism: in their famous 1986 book, Rumelhart, McClelland, and the PDP group (Rumelhart et al., 1986) provided a concrete and detailed account of cognition which was

completely alternative to the symbol manipulation paradigm. According to this alternative view cognition is not the manipulation of symbols according to formal rules, but rather the parallel and distribute processing of sub-symbolic information, that is, the transformation of purely quantitative values (the pattern of activation of groups of units) using other quantitative values (the connection weights linking groups of units) in networks of neuron-like units.

Other fundamental attacks to the classical view of cognition as symbol manipulation came in the early 1990s from behaviour-based robotics and Artificial Life (Brooks, 1990; Parisi et al., 1990). The 'Artificial Life route to Artificial Intelligence' (Steels and Brooks 1994) pointed to the fact that cognitive processes are always 'embodied', 'situated' and (partially) 'distributed' in an organism's environment. They are embodied in that the body and its physical properties are important determinants of the way a given task is solved. They are situated because the constrains provided by the environment can act also as opportunities for the task's solution. And they are partially distributed because they do not happen only inside an organism's head; rather, they crucially depend on the organism's environment which, especially in the human case, includes artefacts and other agents. (For a view of connectionism as part of Artificial Life, in which neural networks control the behaviour of embodied and situated agents see Parisi, 2001).

Finally, another challenge to the symbolic approach to cognition came from dynamical systems theory. Proponents of the dynamical hypothesis argue that cognition should not be accounted for in computational terms, but rather using differential equations and dynamical systems concepts such as equilibrium points, cyclic behaviour, attractors, and bifurcations. More specifically, cognition must be understood by interpreting a cognitive system as a point moving in an abstract multi-dimensional space, and by identifying the trajectories that the system follows in that space and the laws that govern these trajectories (Smith and Thelen, 1993; Port and van Gelder, 1995; van Gelder, 1998; Beer, 2000).

The concepts and tools of connectionism, robotics, and dynamical systems theory opened up several very active areas of research, especially of the synthetic kind. The overall result is that contemporary cognitive science is substantially rethinking its view of cognition. In particular, the fundamental assumption of classical cognitive science that cognition is the manipulation of symbols according to formal rules is being replaced by a view according to which the mechanisms that explain behaviour are non-symbolic or sub-symbolic, and cognition consists in the adaptation of an agent to its environment. Furthermore, this adaptation critically depends on the dynamic interactions between the agent and the environment, which can also include artefacts and other agents (Bechtel et al., 1998; Clark, 2001).

But apart from 'classical' connectionism, which addresses all levels of cognition but without taking into account 'embodiment' and 'situatedness', the new cognitive science has been so far concerned mostly, if not exclusively, with low-level behaviors and capacities, such as perception, learning, sensory-motor coordination, and navigation. The question remains open whether the same broad framework can scale up to explain the higher forms of cognition which characterize human beings (such as problem solving, reasoning, and planning), or if in order to explain characteristic human cognition we must go back to the symbol manipulation paradigm. From the point of view of the new cognitive science the most promising way of addressing this question, we argue, is to consider language not only as a communication system but also as a cognitive tool.

Language as a cognitive tool

The view of language as something that transforms all human cognitive processes dates back as early as the 1930s, with the work of Russian scholar Lev Vygotsky (Vygotsky, 1962; Vygotsky, 1978). According to Vygotsky, the most important moment in child development is that in which the child begins to use language not only as a social communication system but also as a tool for controlling her own actions and cognitive processes. When the child is challenged by a particularly difficult task she is often given help by an adult or a more skilled peer, and this help typically takes a linguistic form. Later on, when the child is facing the same or a similar task all alone, she can rehearse the social linguistic aid which helped her to succeed in the problem. This is called 'private speech', which, according to Vygotsky, plays a fundamental role in the development of all human psychological processes.

The linguistic social aid coming from adults takes several forms. Social language helps a child to learn how to categorize experiences, to focus her attention on important aspects of the environment, to remember useful information, to inhibit non-useful behavior, to divide challenging problems into easier sub-problems and hence to construct a plan for solving complex tasks, and so on. When the child is talking to herself she is just making to herself what others used to do to her, that is, providing all sorts of cognitive aid through linguistic utterances. Once the child has mastered this linguistic self-aid, private speech tends to disappear, but only if one looks at the child from outside. In fact, it is only abbreviated and internalized, becoming inner speech. Hence, most, if not all, of adult human cognitive processes are linguistically mediated, in that they depend on the use of language for oneself.

Recently, the idea of language as a cognitive tool has been given increasing attention within the cognitive-science-oriented philosophy of

mind (Carruthers and Boucher, 1998). For example, Daniel Dennett (Dennett, 1991; Dennett, 1993; Dennett, 1995) has argued that the human mind, including its most striking and hard to explain property, namely consciousness, depends mostly not on innate cognitive abilities, but on the way human plastic brains are substantially 're-programmed' by cultural input coming, principally, through language: "Conscious human minds are more-or-less serial virtual machines implemented -inefficiently - on the parallel hardware that evolution has provided for us" (Dennett, 1991, p. 278).

Andy Clark (Clark, 1997; Clark, 1998; Clark, 2006) has further developed these Dennettian ideas by providing several arguments about how animal-like, embodied, situated, and sub-symbolic cognitive processes can be augmented by the learning and use of linguistic signs. According to Clark, language is not only a communication system, but also a kind of "external artifact whose current adaptive value is partially constituted by its role in re-shaping the kinds of computational space that our biological brains must negotiate in order to solve certain types of problems, or to carry out certain complex problems." (Clark, 1998, p. 163).

Apart of the interesting philosophical ideas of Dennett and Clark, the Vygotskyan view of language as a cognitive tool has recently been raising increasing interest also in empirical cognitive science (see, for example, Gentner and Goldin-Meadow, 2003). Indeed, a growing body of empirical evidence demonstrates the importance of language for a number of cognitive functions including learning (Nazzi & Gopnik, 2001), memory (Gruber & Goschke, 2004), analogy making (Gentner, 2003), cross-modal information exchange (Spelke 2003), problem solving (Diaz & Berk, 1992), abstract reasoning (Thompson et al., 1997), and logico-mathematical abilities (Dehaene et al., 1999).

In our work, we explore and articulate the hypothesis of language as a cognitive tool by the aid of artificial life simulations which use neural networks as models of the nervous system and genetic algorithms as models of evolution by natural selection. Computer simulations can provide fundamental tools in the development of new ideas and in the formulation of theories in that (a) they force the theory to be stated clearly and in full details, (b) they uncontroversially generate the consequences of the assumptions of the theory as the simulation results, and (c) they suggest new ideas and directions of research. In what follows we describe some recent computational models of the use of language as a cognitive aid and of its role in human evolution.

Language and categorization

Basically, organisms respond to sensory inputs by generating motor outputs. The motor output which is generated in response to some

particular sensory input tends to have consequences that increase the individual's survival/reproductive chances. Evolution and learning are processes, respectively at the population and individual level, that result in acquiring the capacity to respond to sensory inputs with the appropriate motor outputs. We model organisms using neural networks and evolution and learning as changes in the networks' connection weights that allow the organism to respond appropriately to sensory input.

If we look at sensory-motor mapping we see that it is not the case that each different sensory input requires a different motor output. *Different* sensory inputs may require the *same* motor output, and different sensory inputs that require the same motor output are said to form 'categories'. (Motor outputs can be 'the same' at some more abstract level than the level of the specific physical movements. An organism can respond to an object with the same action of 'reaching' for the object although in different occasions the specific physical movements of the organism's arm can be different, for example as a function of the arm's starting position.) What are categories in terms of a neural network model of behaviour? To answer this question we have to consider how a simple sensory-motor neural network is structured and functions.

In a neural network some particular sensory input is encoded as some particular activation pattern in the network's input units. This activation pattern elicits another, particular activation pattern at the level of the hidden units, which in turn elicits a particular activation pattern in the output units. The activation pattern appearing in the output units determines the particular movement with which the organism responds to the sensory input. Neural networks learn to respond appropriately to sensory input by modifying their connection weights (either by genetic evolution or through individual learning) so that *different sensory inputs that must be responded to with the same motor output will elicit similar activation patterns in the hidden units*, and *similar sensory inputs that must be responded to with different motor outputs will elicit different activation patterns in the hidden units*. (For an Artificial Life model of this action-based view of categories, see Di Ferdinando and Parisi, 2004.)

We can consider the activation pattern observed in the network's hidden units at any given time as one point in an abstract hyperspace with as many dimensions as the number of hidden units, where the coordinate of the point for each dimension is the activation level of the corresponding unit. Categories are 'clouds' of points in this abstract hyperspace, that is, sets of points elicited by sensory inputs that must be responded to with the same motor output. Different categories are different clouds of points. Good categories are clouds of points that are (a) small (activation patterns that must be responded to with the same motor output are made more similar by the connection weights linking the input units to the hidden units) and (b) distant from each other (activation patterns that

must be responded to with different motor outputs are made more different by these weights). The reason is that effectiveness of the organism's behaviour depends on the goodness of these categories. With good categories the organism will be less likely to respond in different ways to sensory inputs that require the same response, or in the same way to sensory inputs that require different responses.

What are the consequences of the possession of language for an organism's categories? We can model language as a second sensory-motor network which is added to the basic sensory-motor network that we have already described and which underlies the organism's non-linguistic behaviour. We will call the two networks the 'sensory-motor network' and the 'linguistic network', respectively. Like the sensory-motor network, the linguistic network has a layer of sensory input units connected to a layer of hidden units connected to a layer of motor output units. The sensory units of the linguistic network encode linguistic (heard) sound and the motor output units encode phono-articulatory movements that produce linguistic sounds. During the first year of life of the child, the linguistic and the sensory-motor network are not functionally (or perhaps even anatomically) connected and they are used separately. The child uses the sensory-motor network to learn to map non-linguistic sensory inputs from objects and persons into the appropriate motor actions (e.g. reaching for, grasping, and manipulating objects, following another person's gaze, turning towards another person, etc.) and uses the linguistic network to learn to generate phono-articulatory movements that result in sounds corresponding to heard sounds (that is, imitating the linguistic sounds of the particular language spoken in its environment).

At 1 year of age proper language learning begins. The two networks become functionally connected and the child begins to learn the appropriate synaptic weights of the two-way connections linking the hidden units of the sensory-motor network to the hidden units of the linguistic network. What are the appropriate synaptic weights for these connections? These are weights such that a particular sound which is heard by the child, i.e., which is encoded in the sensory units of the linguistic network, will tend to elicit an activation pattern in the hidden units of the sensory-motor network which is similar to the activation pattern elicited by some perceived object or action, and thus in a non-linguistic action which is appropriate to the heard sound. This is language understanding. And, conversely, a particular perceived object or action, which is encoded in the sensory units of the sensory-motor network, will tend to elicit an activation pattern in the hidden units of the linguistic network that result in some appropriate phono-articulatory movements. This is language production.

What are the consequences of this reciprocal functional linking of the sensory-motor network and the linguistic network, i.e., of possessing a language, for the organism's categories? The answer is that

categorization is enhanced by language (Mirolli and Parisi 2005b). When the child hears and understands the language spoken by others, the child's categories tend to become better categories, i.e., smaller and more distant clouds of points in the child's neural network. If the child perceives an object and at the same time she hears the linguistic sound that designates the object in the particular language spoken in her environment, the activation pattern that results in the hidden units of the child's sensory-motor network depends on both the sensory input from the object and the sensory input from the linguistic sound. The consequence is that this activation pattern is more similar to the activation patterns elicited in other occasions by other objects belonging to the same category (that must be responded to with the same action) and more dissimilar to the activation patterns elicited by objects belonging to other categories, compared with the activation pattern observed in an organism without language.

But this is not the whole story. An important characteristic of human language, which distinguishes it from the communication systems of other animals, is that human language is used not only for communicating with others but also for communicating with oneself. Indeed, the use of language for oneself starts as soon as language is acquired, and represents a significant proportion of the child's linguistic production. Empirical studies demonstrate that 3 to 10 year old children use language for themselves 20-60% of the time (Berk, 1994).

As discussed above, the use of language for talking to oneself can be related to the 'language as a cognitive tool' hypothesis: private speech happens as the child discovers that she can exploit the advantages provided by language by talking to herself. Later on, the child can internalize this linguistic self-aid, by just 'thinking' linguistic labels without producing them out aloud. Can this interpretation of private and inner speech be applied to the advantages produced by language on categorization? In order to answer this question we need to model both ways in which humans can talk to themselves: externally, as private speech, or internally, as inner speech.

The simulation of private speech is quite straightforward. The network encounters an object and it responds to the object by producing the sound that designates the object using its linguistic sub-network. Then, the network hears the sound it has just produced and responds, using its sensory-motor sub-network, to the internal representation of the self-produced sound. Inner speech can instead be simulated as follows. When the network perceives an object, it does not produce any sound. Nonetheless, the sight of the object does induce the internal representation of the name of the object in the linguistic hidden units. In inner speech, it is this internal representation of the label associated to the perceived object that influences the non-linguistic response of the network.

As it turns out, the advantage for the network's categories provided by social language, when the network hears linguistic signals produced by somebody else, can be observed even if the organism is all alone and talks to itself. In fact, both self-produced and internally-thought linguistic signals improve sensory-motor internal representations of perceived objects more or less to the same extent as social linguistic input. That is, compared to the representations of the pre-linguistic network, internal representations of objects belonging to the same category are more similar (close) to each other, and those of objects belonging to different categories are more different (distant) to each other (see Mirolli and Parisi, 2006).

Talking to oneself in the evolution of language

Why did language evolve? What was the adaptive function of language? This question is surely of the most importance, if one wants to understand the evolution of language and of man in general. Nonetheless, in the contemporary literature on language evolution there is not much debate on this topic (see, for example, Knight et al., 2000; Christiansen & Kirby, 2003). One reason seems to be the common assumption that the only function of language is communication. As we have discussed in section 2 the 'received view' holds that language is nothing but a very complex and powerful communication system. But once one has acknowledged the importance of language in the development of human cognition one can no more assume that the evolution of language has been driven only by the pressures for better communication. On the contrary, an interesting question immediately rises: when did hominids started to use language for themselves as a cognitive tool?

Generally, there is a tendency to think that language was used by humans to communicate with themselves only when language was already well developed and was sophisticated and syntactically complex; hence, quite recently compared with the first appearance of a proto-language. However, this is not necessarily the case. Even a very simple proto-language, for example, a language made up of single words (or holophrases), may be used to talk to oneself, with advantages for the individual that uses the language in this way. Based on this hypothesis, we have developed another set of simulations in which we studied the effect of talking-to-oneself for the evolution of a simple communication system (Mirolli & Parisi 2005a).

In this simulation a population of artificial organisms (whose behaviour is controlled by neural networks) evolve in a simple world which contains both other organisms and poisonous and edible mushrooms. Organisms must avoid poisonous mushrooms, which decrease an individual's

probability to reproduce, and eat edible ones, which increase individual fitness. Furthermore, organisms can communicate to each other the quality of encountered mushrooms by emitting signals through their linguistic output units. But in order to exploit the advantages provided by communication the population must evolve an appropriate communication system. Since each individual mushroom is different from all other mushrooms belonging to the same category, organisms must evolve the capacity to send similar signals every time they encounter an edible mushroom and another signal when they encounter poisonous mushrooms.

The evolution of such a communication system proves to be quite difficult, especially because in this simulation there is no direct selective pressure for producing the appropriate signals: an individual's reproductive chances depend on the number and quality of mushrooms the individual eats, not on the signals it produces. Indeed, by producing good signals an individual can increase the probability of reproduction of another individual, thus providing a direct advantage for a competitor. The result is that in the standard simulation, in which signals are used only for social communication, a good and stable communication system never evolves.

In another simulation we let organisms use signals not only for sending signals to each other, but also for talking to themselves, as aids to memory. In particular, organisms can hear their self-produced signals and use them in order to remember the information received by other organisms. The results of this second simulation are clear: if organisms can use language not only as a social communication system but also as a cognitive (memory) aid the evolution of language itself is favoured, and this has a positive impact on the organisms' fitness as well. In other words, organisms which can talk to themselves develop a better communication system and reach a higher fitness with respect to organisms which can use signals only for communication. The reason is clear: in order to exploit the advantages provided by using language as a memory aid organisms must produce useful signals, because otherwise they would mislead themselves. In other words, talking to oneself associates a direct individual advantage to producing useful signals, which was not the case in the previous simulation.

Using language as an aid to memory can be advantageous for at least two reasons: (a) delegating the memory function to the linguistic system can leave the sensory-motor system free to process other information useful for acting in the environment while linguistically remembering previous information, and (b) linguistic signals may occupy less space in memory than the sensory-motor information they refer to.

Using language as a cognitive tool may have had a fundamental impact not only on categorization and memory. For example, other neural network simulations have shown that language can improve the *learning*

of categories (Schyns, 1991; Lupyan 2005). Furthermore, the artificial life simulations of Cangelosi and colleagues (Cangelosi and Harnad 2000; Cangelosi et al., 2000) have demonstrated that language can also allow 'symbolic theft', that is, a way of learning useful categories not by direct sensory-motor experience with the world but through cultural transmission mediated by language. And it can be argued that talking to oneself can be useful in many additional ways. It can allow an individual to direct her attention to specific aspects of the environment, to make explicit predictions of future states of the environment, and to explicitly plan future actions (see Parisi and Mirolli 2006).

In as much as the advantages of talking to oneself do not require a complex syntactic language, it is reasonable that the discovery of the cognitive uses of language could have happened quite early in language evolution, in particular before the transition from an holistic proto-language to the full-blown compositional language of modern humans. And this is just what the computational models reviewed here suggest: none of them included any kind of syntax, but just the 'symbolic' capacity to associate 'meanings' (as internal representations of significant experiences) with linguistic labels. Nonetheless, they demonstrated that addressing to oneself even simple linguistic labels can provide important individual advantages. Trying to sort out what could have been the consequences of this early use of language for oneself in the subsequent evolution of language is an interesting topic for future research.

Conclusion

A crucial, but often neglected, characteristic of human language is that language is used not only for communicating with others but also for communicating with oneself, whereas we seem not to have evidence for this type of use of animal communication systems. Talking to oneself, in the form of both private and inner speech, has tremendous consequences for the development of the human mind. Indeed, we have argued that considering the cognitive role of language can provide the missing link for addressing the high-level cognitive capacities which characterize humans within the new, emerging framework which considers cognition as "environmentally embedded, corporeally embodied, and neurally embrained." (van Gelder, 1999, pag. 244). In the present paper we have described some simple computer simulations that show that language can improve one's categories and can be an useful aid to memory, both if it mediates social communication and if it is used to talk to oneself as private or inner speech. But we argue that the use of language for oneself does not improve only categorization and memory, but almost any human cognitive function. Therefore, much more work needs to be done in order to understand the relationships between

language and cognition. And we think that computer simulations will play an important role in our understanding of this fundamental topic.

Acknowledgements

The research presented in this paper has been supported by the ECAGENTS project founded by the Future and Emerging Technologies program (IST-FET) of the European Community under EU R&D contract IST-2003-1940.

References

Bechtel, W.; Abrahamsen, A. & Graham, G. (1998),The Life of Cognitive Science, in William Bechtel & George Graham, ed.,'A companion to cognitive science', Blackwell, Oxford, MA.

Beer, R.D. (2000), 'Dynamical approaches to cognitive science', Trends in Cognitive Sciences 4(3), 91--99.

Berk, L.E. (1994), 'Why children talk to themselves', Scientific American, 78--83.

Brooks, R.A. (1990), 'Elephants Don't Play Chess', Robotics and Autonomous Systems 6, 3--15.

Cangelosi, A.; Greco, A. & Harnad, S. (2000), 'From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories', Connection Science 12(2), 143--162.

Cangelosi, A. & Harnad, S. (2000), 'The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories', Evolution of Communication 4, 117--142.

Carruthers, P. & Boucher, J., ed. (1998), Language and thought: Interdisciplinary themes, Cambridge University Press, Cambridge.

Christiansen, M.H. & Kirby, S., ed. (2003), Language Evolution: The State of the Art, Oxford University Press, Oxford.

Churchland, P.M. (1981), 'Eliminative Materialism and the Propositional Attitudes', Journal of Philosophy 78(2), 67-90.

Clark, A. (1997), Being There: putting brain, body and world together again, Oxford University Press, Oxford.

Clark, A. (1998),Magic words: How language augments human computation, in Peter Carruthers & Jill Boucher, ed.,'Language and thought: Interdisciplinary themes', Cambridge University Press, Cambridge, pp. 162--183.

Clark, A. (2001), Mindware: an introduction to the philosophy of cognitive science, Oxford University Press, Oxford.

Clark, A. (2006), 'Language, embodiment, and the cognitive niche', Trends in Cognitive Sciences 10(8), 370--374.

Dehaene, S.; Spelke, E.; Pinel, P.; Stanescu, R. & Tsivkin, S. (1999), 'Sources of mathematical thinking: Behavioral and brain-imaging evidence', Science 284, 970--974.

Dennett, D.C. (1995), Darwin's Dangerous Idea: Evolution and the Meanings of Life, Simon and Schuster, New York, NY.

Dennett, D.C. (1993), 'Learning and labeling', Mind and Language 8(4), 540--547.

Dennett, D.C. (1991), Consciousness Explained, Little Brown & Co., New York, NY.

Dennett, D.C. (1978), Brainstorms, Bradford Books, Montgomery, Vt.

Diaz, R. & Berk, L.E., ed. (1992), Private speech: From social interaction to self regulation, Erlbaum, New Jersey, NJ.

Di Ferdinando, A. & Parisi, D. (2004), Internal representations of sensory input reflect the motor output with which organisms respond to the input, in A. Carsetti, ed.,'Seeing, thinking and knowing', Kluwer, Dordrecht, pp. 115--141.

Dreyfus, H. (1972), What computers can't do: A critique of artificial reason, Harper & Row, New York.

Fodor, J. (1975), The language of thought, Harvard University Press, Cambridge, MA.
Gallese, V. & Lakoff, G. (2005), 'The Brain's Concepts: The Role of the Sensory-Motor System in Reason and Language', Cognitive Neuropsychology 22, 455--479.
van Gelder, T.J. (1998), 'The dynamical hypothesis in cognitive science', Behavioral and Brain Sciences 21, 1--14.
Gentner, D. (2003),Why we are so smart, in Dendre Gentner & Susan Goldin-Meadow, ed.,'Language in mind', MIT Press, Cambridge, MA, pp. 195--235.
Gentner, D. & Goldin-Meadow, S., ed. (2003), Language in mind, MIT Press, Cambridge, MA.
Gruber, O. & Goschke, T. (2004), 'Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes', Acta Psychologica 115, 105--121.
Knight, C.; Hurford, J.R. & Studdert-Kennedy, M., ed. (2000), The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form, Cambridge University Press, Cambridge.
Lupyan, G. (2005), Carving nature at its joints and carving joints into nature: How labels augment category representations, in Angelo Cangelosi & R. Bugmann, G.a nd Borisyuk, ed.,'Modelling Language, Cognition and Action: Proceedings of the 9th Neural Computation and Psychology Workshop', World Scientific, Singapore, pp. 87--96.
Mirolli, M. & Parisi, D. (2005a), 'How can we explain the emergence of a language which benefits the hearer but not the speaker?', Connection Science 17(3-4), 325--341.
Mirolli, M. & Parisi, D. (2005b), Language as an aid to categorization: A neural network model of early language acquisition, in Angelo Cangelosi; Guido Bugmann & Roman Borisyuk, ed.,'Modelling language, cognition and action: Proceedings of the 9th Neural Computation and Psychology Workshop', World Scientific, Singapore, pp. 97--106.
Mirolli, M. & Parisi, D. (2006), Talking to oneself as a selective pressure for the emergence of language, in Angelo Cangelosi, Andrew Smith & Kenneth Smith, ed.,'The Evolution of Language: Proceedings of the 6th International Conference on the Evolution of Language', World Scientific Publishing, , pp. 214--221.
Nazzi, T. & Gopnik, A. (2001), 'Linguistic and cognitive abilities in infancy: When does language become a tool for categorization?', Cognition 80, 303--312.
Newell, A. & Simon, H. (1976), 'Computer science as empirical inquiry: Symbols and search', Communications of the ACM 19, 113--126.
Parisi, D. (2001), Neural networks and Artificial Life, in David Baltimore, Renato Dulbecco, Francois Jacob & Rita Levi-Montalcini, eds., ‘Frontiers of Life’ Volume 4, Academic Press, San Diego.
Parisi, D.; Cecconi, F. & Nolfi, S. (1990), 'Econets: Neural networks that learn in an environment', Network 1, 149--168.
Parisi, D. & Mirolli, M. (2006),The emergence of language: How to simulate it, in C. Lyon; C. Nehaniv & A. Cangelosi, ed.,'Emergence of Communication and Language', Springer Verlag, Berlin.
Port, R.F. & van Gelder, T., ed. (1995), Mind as Motion, MIT Press, Cambridge, MA.
Putnam, H. (1963),Brains and Behavior, in R. Butler, ed.,'Analytical Philosophy', Basil Blackwell, Oxford.
Rumelhart, D., E., McClelland, J.L., and the PDP research group (1986), Parallel distributed processing: Explorations in the microstructure of cognition, MIT Press, Cambridge, MA.
Schyns, P.G. (1991), 'A Modular Neural Network Model of Concept Acquisition', Cognitive Science 15(4), 461--508.
Searle, J.R. (1980), 'Minds, brains and programs', Behavioral and Brain Sciences 3, 417--458.
Smith, L.B. & Thelen, E., ed. (1993), A Dynamic Systems Approach to Development, MIT Press, Cambridge, MA.
Spelke, E. (2003),What makes us smart? Core knowledge and natural language, in Dendre Gentner & Susan Goldin-Meadow, ed.,'Language in mind', MIT Press, Cambridge, MA, pp. 277--311.
Steels, L. & Brooks, R., ed. (1994), The artificial life route to artificial intelligence: Building Situated Embodied Agents, Lawrence Erlbaum Ass., New Haven.
Thompson, R.K.R.; Oden, D.L. & Boysen, S.T. (1997), 'Language-naive chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task', Journal of Experimental Psychology: Animal Behavior Processses 23, 31--43.

Vygotsky, L.S. (1978), Mind in society, Harvard University Press, Cambridge, MA.
Vygotsky, L.S. (1962), Thought and language, MIT Press, Cambridge, MA.
Waxman, S. & Markow, D. (1995), 'Words as invitations to form Categories: Evidence from 12 to 13-month-old infants', Cognitive Psychology 29(3), 257--302.

IS EVOLUTION ALGORITHMIC?

Marcin Miłkowski

Institute of Philosophy and Sociology

Polish Academy of Sciences

+48226351350

marcin.milkowski@ifispan.waw.pl

Abstract. In *Darwin's Dangerous Idea*, Daniel Dennett claims that evolution is algorithmic. On Dennett's analysis, evolutionary processes are trivially algorithmic because he assumes that all natural processes are algorithmic. I will argue that there are more robust ways to understand algorithmic processes which make the claim that evolution is algorithmic empirical and not conceptual. While laws of nature can be seen as compression algorithms of information about the world, it does not follow logically that they are implemented as algorithms by physical processes. For that to be true, the processes have to be part of computational systems. The basic difference between mere simulation and real computing is having proper causal structure. I will show what kind of requirements this poses for natural evolutionary processes if they are to be computational.

Daniel Dennett made a claim that evolution is algorithmic (Dennett 1995: 60). Several authors objected that on Dennett's analysis, evolutionary processes could be trivially algorithmic because he assumes that all natural processes are algorithmic (Fodor 1996: 253, Ahouse 1998: 361-363; cf. Dennett 1995: 59). This objection is misleading if all natural processes aren't algorithmic in the sense Dennett wants evolution to be algorithmic. It isn't at all trivial that evolution is algorithmic if all physical processes aren't computational. Pancomputationalism, or universal computationalism is the claim that all physical processes are computational but this, on my strict criteria of computing, will turn out false (for other criticisms of universal computationalism, see Piccinini 2007). The question is how to understand "algorithmic". What would make the claim about the evolution true?
There are processes effectively describable by computations ("algorithmic" in Gregory Chaitin's sense of algorithmic information theory, cf. Chaitin 1975), and processes that realize digital computations. (I am ignoring analog computation here for two reasons. There is no standard analog computation algorithm theory, and the claim I am trying to evaluate is far more controversial when it refers to digital computation.) In what sense are evolutionary processes algorithmic?
All natural processes are algorithmically describable. In this regard, Dennett was right to say they are algorithmic. This is trivial, given the standard algorithmic information theory, though die-hard materialists would probably disagree (see Mahner and Bunge 1997). Yet it's highly controversial that any biological or evolutionary processes are computational. While laws of nature can be seen as compression algorithms of information about the world, it doesn't follow logically that they are implemented as algorithms by physical processes. For that to be true, the processes have to be part of computational systems. The basic difference between mere simulation and real computing is having proper causal structure (Scheutz 2002). Dennett is probably right if he means the weaker claim (evolution can be simulated), and there aren't many reasons to think he's right if he means the stronger, computational claim. That's why Gould (1997) could be right but not because of the reasons that he referred to, as I will show.
Dennett has to defend the specific claim about evolution independently from any claims about all processes, if he didn't mean the first one to be a trivial consequence. However, Dennett's definition of algorithmic processes cannot account for the simulation/computing distinction. Algorithmic processes, according to Dennett, have three features:

1. substrate neutrality
2. underlying mindlessness
3. guaranteed results

The problem is that all functionally describable processes share these features. For example, the process of opening a can be realized using a simple, hand-operated device or engine-driven device, so it's substrate neutral. Opening the can isn't rocket science, either, and it has guaranteed results (*ceteris paribus*). All computational processes are functionally specifiable processes but they have more distinguishing features. A hand-operated can opener isn't a computer, after all. If we fail to see that, we will follow Putnam-Searle fallacy of ascribing any computation to any process: Nothing would disallow ascribing realization of Wordstar program to Searle's notorious wall (Searle 1992: 207-208). Moreover, any disjunction of states of the processes can be thought to realize a computational process, and using arbitrary disjunctions on sufficiently complicated systems, we could ascribe them any possible computation. These results aren't only counterintuitive; they follow from a definitional *fiat* that Searle and Putnam made: They understand computation as a purely syntactic (or formal) object.
Searle argues "syntax is not intrinsic to physics" (Searle 1992: 210). If Searle means that physics isn't linguistics, he's right. Nevertheless, he's wrong to treat algorithms as purely formal syntactic objects. This formalism, if consistent, would make him deny the reality of all mathematical properties ascribed by physicists. It isn't the fact that all mathematical properties ascribed in physics are observer-relative. Properties of computer programs are just the same as the rest. They should be ascribed in the same way science generally ascribes mathematical values to objects.

Computational systems

Contrary to such broad realization concepts, stricter criteria have been recently proposed (see Miłkowski 2006, Scheutz 2001). The list that I'll present is preliminary, and I'll supplement it with a general requirement connected with functional systems as such.
Computational systems are functional systems. There are at least three ways to analyze these systems, according to the notion of function used. First, there is Cummins' notion of function as causal role of a part of the system (Cummins 1975). Second, there are history-based notions, such as defined by Wright (1973) or Millikan (1984). Third, there's a design-based notion of function, as defined by Ulrich Krohs (2004). Cummins' notion is very broad, and makes any causal role a function – for example, the function of the trash can lid is making noise in the middle of the night. According to history-based notions, the function is the cause (a reason) that a thing that has it exists, so prototypes have no functions. Krohs' notion needs a little more explanation.

Krohs suggests that all functional systems have design that specifies system parts in terms of part types. For example, if I want to assemble my IKEA table, I read the manual (the design specs) that specifies the screws, but not as individuals with proper names or located in space-time, but as types. In case of biological systems, the genotype specifies the design. Human-made functional artifacts have parts selected as types by humans; other functional systems are selected by other mechanisms (natural selection seems the most obvious one). This notion has an obvious advantage: the design stance descriptions are literally descriptions based on ascribing design ascriptions. For this very reason, this notion seems appropriate for analyzing Dennett's claims: The design stance would turn out to be based on the notion of design. The task of re-engineering of artifacts and biological systems could be then reformulated as the task of rediscovering their design: their specification in terms of part types and relation of these parts.

Based on these three kinds of notions, three types of functional systems could be defined. The choice of the notion has deeper consequences – probably anything would be a functional system in Cummins' terms but not according to other notions. Prototype systems won't be functional in Millikan's terms, and systems without type-level specifiable parts won't be functional if we accept Krohs' criteria. This means, for example, that dissipative systems which are easily described as wholes in terms of types aren't functional: Their individual parts cannot be picked out using any type-level description—there aren't type-selection mechanisms that would allow for functional ascriptions. Just because dissipative systems are physical systems but not functional systems, they cannot be computational systems, and universal computationalism is false. At the same time, universal computationalism goes hand in hand with Cummins' like functions because parts of dissipative systems could be ascribed causal roles.

The computational description should offer new predictions or explanations. If it isn't the case, the computational description of a given system is redundant, and it's safe to say that the system isn't computational. For example, working of a can opener can be described without stipulating any computation; the can opener doesn't need to process any information about the can to open it (at least that's how today's can openers work). This is just a general rule of stipulating higher-level properties; if lower-level properties are sufficient to predict or explain the behavior and innards of a system, it makes little sense to ascribe higher-lever properties (e.g., it's just as useless to ascribe intentional properties to a lawn). The rule can be spelled out more precisely in terms of Chaitin's algorithmic information theory: the computational description must be simpler than the lower-lever description (a general causal-role level description) and offer epistemic advantages such as new predictions and explanations. The simplicity

boils down to the length of the description (it's equivalent to the compression ratio of the new description compared to the old one). This requirement conflicts with trivial versions of universal computationalism. If universal computationalism could offer new insights for every single physical object, then it would be compatible with the requirement.

The description must be applied consistently for all events in the physical system. We can easily imagine "cheating": devising ascription rules that are far more complex than the system being described, picking out arbitrary disjunctions of states, and so on. This requirement is obvious but notorious "proofs" that any system can perform any computation (Putnam 1987) are so widespread that we should be explicit about the ascription rules. Anyway, that's how natural sciences ascribe mathematical properties, so it shouldn't be controversial.

Ascriptions of sequences of computational states to the system must reflect its causal history. This is just an extension of the consistency requirement into causality. Arbitrary disjunctions of states won't count as causal history so they cannot be described as real computation. This also disallows universal computationalism based only on formal tricks.

The system realizing computations is relatively isolated from the environment. Only functional systems are computational systems, and a system is functional only when it has identifiable boundaries. The boundaries could be blurry but they must delineate the system from the environment. I would define system boundaries in terms of causal relation frequency: causal relations are more frequent inside the system than outside. Even input-output causal relations with a computational system don't make inputs automatically inner values: input relations can obtain with many different objects, which mean that they will be less frequent than real inner relations. If input relations are always connected with the same object or process, this process is a part of the system. This way my delineation criteria help to understand why the notion of extended mind seems intuitive in some cases: it's intuitive only when a remote part of the cognitive system is in fact its subsystem.

It could be argued that some other physical property (other than causal relation relative frequency) should be used to define system boundaries. For example, those who oppose extended mind theories could claim that system boundaries should be spelled out in functional type-level terms of system organization. This kind of system boundary definition is acceptable, as well. What is important is the fact that system boundaries should be definable not only on a computational level of description. Note that arbitrary process state disjunctions nor Searle's wall cannot be clearly delineated on any other level than computational. This poses also another difficulty for universal computationalists because it requires them to show that all physical objects are parts of relatively isolated physical systems.

As I already mentioned, computational systems normally have input states. On the one hand, input data can be internal part of the algorithm the system is implementing. The output data, on the other hand, must be always present. Input and output states should be specifiable, as before, not only on a computational level of description. Note that Searle's wall has no clear input states: there is no wall equivalent of the keyboard nor of the display. Searle hasn't shown any clear way to pick output states nor input states from the set of all states of his wall. There is no computation without output states. Any object can be ascribed a trivially simple output value: Any property could be said to encode it. But this property must be causally related to the input value. So while most objects could be assigned trivial identity transformation (the output property is the input property), non-trivial computations are harder to show.
The input/output requirement is a result of the standard computation definition in terms of recursive functions (as normally Church-Turing thesis is understood). The whole computational process in the system must have a description in terms of recursive function (or any other equivalent model of computation, like Markov strings, Turing machines, register machines etc.). Computational ascription is a real ascription only on the condition that we know what computation we are ascribing. The computation should be spelled out precisely as code or—at least—as pseudo-code.
To sum up, there are several criteria of computational ascriptions:

1. computational description simplicity, predictive and explanatory value
2. description consistency for all processes in the system; causal determination of ascriptions
3. relative system isolation and non-computational boundaries
4. availability of output states connected causally with input states (if any)
5. specification of code-level description

The concept of function realization, which subsumes the realization of computations, depends on how broadly we understand functions. On the design-based notion, pancomputationalism is false. Is Dennett's computational claim false as well?

Evolution as computation

The above top-down analysis of computational systems shows that if there's a real computational level of description of natural evolutionary processes, this cannot be the only level of their description. Could a

computational description of evolutionary processes fulfill the abovementioned criteria?

The computational description will be simpler than the lower-level physical description. Its explanatory value remains, however, at best controversial: It isn't at all clear what it would explain. Origin of biological information as selected from the chaos? Or the way natural selection works? The predictive value isn't clear neither. Whereas the general algorithm of evolution could predict the way natural selection works in every case, it would probably be highly dependent on the complete knowledge of environmental constraints and details of evolution units being selected. It isn't clear that these predictions wouldn't be available in the modern neo-Darwinian Synthesis. For the sake of argument, let us suppose that we would gain an insight into how evolution, or Mother Nature in Dennett's terms, processes information about replicators and interactors (Brandon 1998).

We would apply the description consistently, based on causal relations. Therefore, we assume that consistency requirement would be fulfilled.

Evolutionary processes are probably relatively easy to single out from other processes (say, geological) but it isn't obvious whether the most relevant elements of these processes have any function in the evolutionary computational systems. Are evolutionary processes like dissipative processes? According to the more robust, design-based notion of function, physical processes can implement algorithms but not all kinds of physical processes are computational: Those that form non-linear and non-aggregative systems that strongly depend on token-only properties like space-time localization cannot have functional elements. If it could be shown that the way evolutionary processes run depends only on their localization (or any purely token-level property), Dennett's claim would be false. At the first glance, this is what Gould wants us to believe:

> Crank your algorithm of natural selection to your heart's content, and you cannot grind out the contingent patterns built during the earth's geological history. You will get predictable pieces here and there (convergent evolution of wings in flying creatures), but you will also encounter too much randomness from a plethora of sources, too many additional principles from within biological theory, and too many unpredictable impacts from environmental histories beyond biology (including those occasional meteors)—all showing that the theory of natural selection must work in concert with several other principles of change to explain the observed pattern of evolution (Gould 1997).

Gould thinks that contingency – responsible for all variability of the population – plays such an important role in natural selection that its algorithm cannot be realistic without considering this contingency. However, contingencies, or initial state of the environment fed into the computational evolutionary process can be treated in two ways: first, they can be described using lossy compression, and second, simply be input into the more complex computational system. Both ways are compatible

with a notion of algorithm realization. What Gould hasn't shown is that these contingencies would make the natural selection algorithm computationally intractable because of the combinatorial explosion.

To answer the question whether it would be computationally tractable, we need the code. What should this code compute? A general natural selection problem or a specific selection problem? According to Gould, we could produce an algorithm for convergent evolution, so this could be a third possibility.

Let's start with the first possibility: a general natural selection algorithm. The fitness of units being selected naturally shows that the solution of the problem of adapting to environment was effectively solved. I would propose that the evolutionary algorithm solves the problem of adaptation, and this fitness or the adapted population could be thought of as output value of the computation. Maybe the interaction with environment could produce the input of this algorithm.

Some hints about what evolutionary computational systems are and what kind of computations they realize can be found in computer science. Research on artificial life or evolutionary algorithms seems to suggest that though there are emergent properties and strong context-dependence of properties, at the same time objects are computational (Crutchfield and Mitchell 1995). It's an empirical question whether natural evolutionary processes are like dissipative systems or rather like Artificial Life.

Evolutionary algorithms are heuristic search algorithms modeled after natural processes (Michalewicz 1996). They involve generating and mutating a population of artificial organisms, and testing them according to a fitness function. The fitness is assessed based on how well the given organism finds a solution to a problem. There are various types of evolutionary algorithms, and not all properties and types of evolutionary algorithms are now known. Most likely, existing evolutionary algorithms are only a small subclass of all possible evolutionary algorithms. Compared to natural processes, artificial ones are less complex but can serve as a starting point for evaluating Dennett's claim.

The problem with the code inspired by the research on evolutionary algorithms is that it cannot be adopted directly. The overall structure of evolutionary algorithms is as follows:

```
procedure evolution
begin
   t=0
   determine_starting_P(t)
   (* P(t) - Population P at time t *)
   final_condition=(evaluate P(t) >threshold)
   while not (final_condition) do
   begin
      t=t+1
      select P(t) (* from P(t-1) *)
      modify P(t)
      final_condition=(evaluate P(t) >threshold)
```

```
   end
end
```

The problem is that this algorithm is based on the *evaluate()* function. This function is however encoded by the programmer, not discovered by the algorithm itself. No general algorithm of fitness assessment seems viable, though here various flavors of adaptationism could have their say. In nature, the encoding of the fitness function is unknown. The fitness landscape is not represented numerically in reality. Therefore, the straightforward application of such algorithms results in a very unsatisfactory code:

```
procedure evolution
begin
t=0
fix starting P(t)
while not (the_end_of_the_world) do
         begin
         t=t+1
         select_naturally_in_environment P(t)
         modify P(t)
         end
end
```

This code cannot possibly fulfill the requirement of explanatory value. It isn't giving any new predictions, and seems only a trivial and redundant reformulation of known causal mechanisms of natural selection.
Replacing the explicit fitness function representation with ways of discovering the environmental constraints might be one of the ways out of this problem. Natural evolutionary computation cannot represent fitness functions that merely make it easier to humans to simulate the causal relations between populations and their environment.
However, it might be argued that general code structure is, in principle, always sketchy and trivial. What we need to find is a specific code for specific evolutionary processes. It might be inspired by current research on evolutionary algorithms or not. In other ways, we should use the bottom-up method to try to find the code.
So I turn to the second possibility, namely to the code computing a specific natural selection process. The arms race between bacteria and antibiotics has been simulated “in silico” by various researches. It's even possible to use “in silico” models to discover new drugs (Gray and Keck 1999). These models start with bacteria genome sequences and proteins expressed by genomes, and knowledge about genes that are crucial for the survival of the bacteria. Such massive data can be then used to predict if certain bacteria or their mutation can survive at all. The general structure of the first algorithm could be used, with appropriate substitutions. The *evaluate* function would test if the bacteria survives when certain proteins are destroyed; it wouldn't require a separate representation.

Does it allow us to say that computational simulation of bacteria vs. antibiotics is really algorithm realization? Not at all. Current "in silico" methods often use data gathered from parallel *in vivo* experiments because scientists still don't know what's being ignored in the simulation. The simulation doesn't include all the causal-functional details, and some of them probably should be disregarded for the sake of simplicity in many cases. Nevertheless, the detailed "in silico" experiment could, in principle, fulfill all criteria of computation realization.
What about convergent evolution? Gould clearly sees that there is a regular pattern in the evolution of the wing in many species. The evolution of the wing could be regarded as an engineering problem, requiring optimization methods. Evolutionary algorithms are used for airplane wing optimization (Keane and Petruzzelli 2000), and a recombination of the existing wing optimization with special organic wing requirements would give us an algorithm for selecting the wing in many different species. The fitness function would be based on aerodynamic features and general engineering principles.
Specific algorithms aren't prone to the problem of how to evaluate the fitness generally. Yet all three sketched algorithms share another disadvantage. They are simulation algorithms rather than algorithms of natural information processing. It cannot be proved that there isn't any other algorithm for natural selection in play, as the problem of existence of any algorithm *for* something is itself not decidable in general: the only way to prove that there is an algorithm for something is to show it.
Biologists, even computational biologists, generally don't seek for computational structure in natural selection. They either build artificial computational systems with biological parts or simulate biological processes. This could mean that, after all, Dennett was right that there are algorithmic processes *in vivo*. But only in Chaitin's sense of the term. Dennett's definition of algorithmic process is redundant then, and reduces easily to the Chaitin's technical term. That's why it doesn't account for simulation/computation distinction. Simulation can sometimes produce a genuine article, for example in a simulated theorem prover. However, in case of evolutionary algorithms it's only a description of evolutionary processes that they produce, and not adapted populations. Moreover, a description shouldn't be confused with what it describes.

Algorithm, natural law, real pattern?

So maybe Dennett's claim isn't about computational powers of evolution but rather about real, multiple-realizable patterns of evolution. These patterns are algorithmic in Chaitin's sense: they aren't stochastic, and there is a way to see regularity in them. After all, the biological data is not all but noise.

A vague usage of “algorithm” is often found in biological papers. Manfred Eigen writes “Our task is to find an algorithm, a natural law that leads to the origin of information” (Eigen 1992, 12). As Mayr notes, biologists often use “models”, “algorithms”, “theories”, “conjectures” interchangeably (Mayr 1997). The problem with this usage, which is roughly compatible with Chaitin’s notion of the algorithmic, is that nature doesn’t realize algorithms for describing the natural processes, and Chaitin’s algorithms are algorithms for describing sequences of information. Therefore, while a process could be algorithmic in Chaitin’s sense, there could be no algorithm that it implements. The implementation could be found somewhere else, for example in a human observer. In other words, evolutionary processes are algorithmic in this sense but aren’t necessarily doing any computations whatsoever.

This notion of the algorithmic is completely compatible with multiple realizability and substrate neutrality. In short, all functionally specifiable processes are algorithmic in Chaitin’s sense, and algorithmic descriptions could be re-used (as non-token level descriptions) to refer to potentially many objects, also made of some other stuff. This leads to a conclusion that multiple realizability of natural selection could be still maintained but natural selection would be no algorithm, only a process in functional systems. But would anyone try to argue with it?

The weaker reading of Dennett’s claim would still face resistance. In principle, functional systems could include cultural processes – as suggested by memetics – or various units of natural selection. There are materialists who claim that it is stuff that matters, and they would object that natural selection isn’t multiple realizable nor functional (Mahner and Bunge 2001). However, this notion of algorithmic processes doesn’t involve any implementation of formal properties or “syntax” by natural selection and this is the premise on which their objection depends, just like in the case of Searle.

Yes, natural selection processes are lawful and not stochastic. They are real patterns. Yes, this is trivial. And it’s much more interesting to see their functional structure in specific cases rather than to say they’re generally algorithmic.

References

Ahouse, J. C., 1998, “The Tragedy of a priori Selectionism: Dennett and Gould on Adaptationism,” *Biology & Philosophy*, 13, 359-391.

Brandon, R., 1998, “The Levels of Selection: A Hierarchy of Interactors,” in: Hull D. and M. Ruse, *The Philosophy of Biology*, Oxford University Press, p. 176-197.

Chaitin, G. J., 1975, “Randomness and Mathematical Proof,” *Scientific American,* 232, No. 5, 47-52.

Crutchfield, J. P. and M. Mitchell, "The Evolution of Emergent Computation", *Proceedings of the National Academy of Sciences, USA* 92:23 (1995), 10742-10746.

Cummins, R., 1975, “Functional Analysis,” *The Journal of Philosophy*, 72, No. 20, 741-765.

Dennett, D., 1995, “Darwin’s Dangerous Idea: Evolution and the Meanings of Life,” Simon & Schuster, Nowy Jork.

Eigen, M., 1992. *Steps toward Life: A Perspective on Evolution*. Oxford: Oxford University Press.

Fodor, J., 1996, “Deconstructing Dennett’s Darwin,” *Mind & Language*, 11, No. 3, 246-262.

Gould, S. J., 1997, "Evolution: The Pleasures of Pluralism," *The New York Review of Books*, June 26, 1997, 47-52.

Grey, C.P. and W. Keck, “Bacterial targets and antibiotics: genome-based drug discovery,”, *Cellular and Molecular Life Sciences*, 56 (1999), 779-787.

Keane, A. J. and N. Petruzzelli, "Aircraft wing design using GA-based multi-level strategies", pp. A00-40171 AIAA-2000-4937 in *Proceedings of the 8th AIAA/USAF/NASSA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, A.I.A.A., Long Beach (2000).

Krohs, U., 2004, *Eine Theorie biologischer Theorien*, Springer Verlag, Berlin.

Mahner, M. and M. Bunge, 2001, “Function and Functionalism: A Synthetic Perspective,” *Philosophy of Science*, Vol. 68, No. 1, 75-94.

Mayr, E., 1997, “This is Biology. The Science of the Living World,” Belknap Press.

Michalewicz, Z., 1996, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer Verlag.

Miłkowski, M., 2006, “Is computationalism trivial?,” In Gordana Dodig Crnkovic and Susan Stuart (eds.), *Computing, Philosophy, and Cognitive Science*, Cambridge Scholars Press (forthcoming).

Millikan, R.G., 1984, *Language, Thought, and Other Biological Categories. New Foundations for Realism*, Cambridge. MA: MIT Press.

Piccinini, G., 2007 “Computational Modelling vs. Computational Explanation”, *The Australasian Journal of Philosophy* (forthcoming).

Putnam, H., 1987, *Representation and Reality*. Cambridge, MA: MIT Press.

Scheutz, M., 2001, “Computational versus Causal Complexity,” *Minds And Machines*, *11*, 543-566.

Scheutz, M., 2002, “Philosophical Issues about Computation”. In *Encyclopedia of Cognitive Science*, London, UK. MacMillan Publishers.

Searle, J., 1992, “Rediscovery of Mind,” MIT Press, Cambridge (Mass).

Wright, L., 1973, “Functions,” *The Philosophical Review*, 82, No. 2, 139-168.

ATTENTION, INTENTION AND THE NATURE OF BELIEVABILITY

PETER WALLIS,
Department of Computer Science,
Sheffield University,
+44 (0)114 222 1800
pwallis@acm.org

Abstract:

A scene can be beautiful; music and games can be engaging, but it takes characters to have a good story. Many have argued that computational models of emotion, and of recognising and presenting emotion, are the key to building believable synthetic characters. This paper looks at what makes an agent believable, and concludes that what matters is intentionality. That is, a character is believable if it behaves in accordance with "the Intentional Stance." Emotions and other mental attitudes play a part in "explaining why" characters do what they do, but emotions are just part of our folk psychological understanding of human behaviour. The paper ends with a discussion of the role of qualia in human evolution.

Introduction

In 2000 I was responsible for the dialogue system for a well-funded project to develop an embodied conversational agent, or ECA, that would act as an interface to a range of data sources and software applications.

The motivation for the project was primarily to explore the issues in developing practical ECA systems, and to distinguish the core technology from the bells and whistles. In preliminary experiments my team looked at the simplified problem of using an ECA to provide access to a database of corporate vehicles and their use. Having many years experience with AI techniques - specifically in natural language processing - the aim was not to have a natural language system that *is* an interface to a database, but to have a software agent that knew *about* the database. Rather than providing stone-wall answers, the system should negotiate a shared ontology, interpreting the user's utterances in the context of the data available, and more importantly telling the user what queries made sense.

The key finding was that politeness was more important than correct understanding [Wallis et al, 2001]. People make mistakes all the time but have a repair mechanism which means the other person doesn't notice, or at least doesn't care. The aim became to create an ECA that could play along with the human. If we could keep the human thinking of the ECA as a person, then the human wouldn't notice the mistakes. If we could make the agent believable, it would simplify the NLP problem by leaving space for errors in things like ASR word recognition and out of domain utterances. Solving the problems of believable characters would by-pass some of the classic issues in NLP.

The conclusion I have come to is that people are very good at "reading minds" - we have a theory of other minds that works well - and hence a believable agent is one who's mind we can read.

This paper starts with a discussion of three approaches to developing synthetic characters and goes on to give evidence in favour of the idea that what matters is folk psychology. The paper goes on to give an algorithm for implementing it, and finishes with a proposal for why we actually *feel* emotion rather than simply acting as if we do.

Human-like Computer Characters

There has been considerable interest in believability - particularly from the games and entertainment industry. Much of the published work on human-like characters in virtual worlds has focused on mimicking human behaviour, but it is not clear that verisimilitude is a necessary part of virtual characters. A quick tour of film provides a menagerie of agents that are simply not human-like, but which are believable. From cartoons to Kafka, the audience is expected to play along with the idea that sponges can talk, and that a person can slowly change into a giant bug.

Not only does it seem to be unnecessary for fictional characters to be just like people, similarity can be a negative attribute. Mori introduced idea of an "the uncanny valley" where synthetic human-like agents are so close to perfect that they appear "uncanny" and actually alienate human interactants (see Ishiguro [Imai et al]). It seems there is some essence of human nature that we recognise. What is it about a character in a play or story that enables us to engage with them, and build a sympathy and understanding of their actions? A popular speculation is that believable agents must express and recognise human emotions.

Emotions

Bates [Bates 94] posited that emotion was key to believability. When children play, the stories told often centre around characters being upset, liking and disliking, and the means of making them happy. Perhaps that is what makes an imaginary character believable; the expression of emotion. Damasio in his book "Descartes Error" [Damasio] points out that there is more to normal human decision making than cold rationality and agues that we need a better understanding of emotion if we are to understand the way humans operate in society. Perhaps synthetic characters also need to recognise and express appropriate emotional responses. The expression and experience of emotion is the focus of considerable research and the HUMAINE Network of Excellence [humaine] is a recent European effort to bring together those interested in the role of emotion in human-computer interaction. The aim is to "... lay the foundations for European development of systems that can register, model and/or influence human emotional and emotion-related states and processes - 'emotion-oriented systems'."

In part this interest in emotion might be driven by our need for identity. Emotion seems to have become the new 'essence' of human nature, distinguishing us from the animals and others. In the past it was our ability to do rational thought but when computers are popularly conceived to be rational, the popular conception of what it is to be human has moved on. From Mr Spock in 1960s Star Trek, to Commander Data in the modern version, what distinguishes the human from the other is the ability to experience emotion. It seems obvious computers cannot experience emotion in the same way people do. One can simulate an aircraft flying through the air on a computer, but a simulation of someone doing mathematical puzzles is actually doing mathematical puzzles. Although a thinking machine is genuinely thinking, it seems obvious that a machine that expresses emotions is only simulating it.

Although the experience of emotions might be key to our sense of identity as humans, this does not mean it is key to a good story. Once again from the history of film we find examples. The hero in a Western is usually a John Wayne-like character that has troubles expressing his emotions. This is not to say the characters are simple; Westerns have a strong notion of suspense and indeed can produce strong emotion in the audience [Ephron]. The point is that the emotion is not expressed by the believable character at the centre of the story. So if emotion, in its raw form, is not the essence of believable action, what are the necessary and sufficient attributes of a suitable character?

The Theatre

The nature of engaging and believable characters is an issue as old as theatre itself. Note that 'engaging' and 'believable' are not the same thing, but perhaps two sides to the same coin. Things can be engaging but not believable - snakes and ladders for example - or believable but not engaging - Tolstoy's "War and Peace" perhaps. If something is not engaging then people get bored. If they are not believable then the audience suffers from what Coleridge referred to as " a failure of our willing suspension of disbelief". Coleridge's point was that the theatre audience participate in a joint delusion, and the aim of the actors is to maintain the audience's willingness to play along. We have all experienced that jolt when something happens in a film where we think "Oh that wouldn't happen!" The flow of the narrative is interrupted and one is reminded that one is in an audience, and these are actors. In work on a virtual assistants, this 'jolt' is the phenomena we needed to avoid.

Breaking the audience's willing suspension of disbelief is something to avoid, but is there any guidance on how to to do that? The answer seems to be, no, not really. Today authors tend to avoid the deus ex machina solution to impossible situations in which (a) god descends from the machinery to make everything all right and the play has a happy ending. But even that is not an absolute. There is a principle however (personal communication with professional script writers) that anything goes, as long as the rules of the virtual world are introduced well before the conflict or pinnacle [Meadows] of the story is reached. The aim of a script writer is to set up a world for which the rules are known. Character and plot development are then executed within that world. As mentioned above, the setting can be very strange; as long as the characters involved have that certain something.

The anterior paracingulate cortex

It seems people have dedicated hardware (wetware) for dealing with other people. When someone thinks they are dealing with another person, their anterior paracingulate cortex becomes more active [Gallagher et al]. We humans are, it seems, hard-wired to deal with other people and in particular to predict the behaviour of others. What is more, our ability to predict is something we use every day and nearly all the time. Walking down the street, we interpret someone striding along as going somewhere; someone looking about is a tourist. Given we do have such a mechanism, it would be reasonable to assume that situations that exercise that mechanism would appeal to us. In the same we enjoy sports which exercise the body, and do sudoku and crosswords to exercise the mind, we watch theatre to exercise our social skills. The proposal is that believable characters are ones that behave in accordance with our mind-reading skills. That is, a character in a play, in a computer game, or assisting us with a database query, will be a believable character while it behaves in accordance with our expectations. This is not to say that the character has to be completely predictable, but that our anterior paracingulate cortex dictates the range of behaviours that are acceptable.

	not engaging	engaging
believable	Tolstoy's "War and Peace"	a Flight Simulator
not believable	Sylvester Stallone movies	Snakes & Ladders

Figure 1: Engaging versus Believable.

So what sort of behaviour are we talking about? Dennett's "intentional stance" [Dennett] explicitly describes the type of thinking we humans do about other rational agents. When we think about physical systems like the game of snooker, we can use what Dennett calls the 'physical stance' to predict what will happen. We reason in terms of cause and effect, and can predict that the blue ball will go in the corner pocket. When reasoning about more complex systems such as alarm clocks, we use a 'design stance.' That is we know what it is designed to do and, without looking at the internals, can predict that the alarm will ring at 6.30 am. When it doesn't, we say 'Oh it is broken' - it is not doing what it is designed to do. With very complex mechanisms, such as human beings or chess-playing computers, we humans use the 'intentional stance'. That is we ascribe the system mental attitudes such as wanting X, loving Y, and intending to do Z. That is, we use 'folk psychology' and assume that rational agents will do what they believe is in their interests. Note that we are sometimes wrong, and psychologists get very interested in the cases where people do not do what (we perceive as) being in their

interests. But in the same way as folk physics is useful even if it is wrong - snooker balls do not slow down of their own accord - folk psychology is usually right enough to provide effective predictions. Usefulness and truth are separate issues in one's Theory of (other) Minds. The status of folk psychology is not clear and Ravenscroft [ref] gives pointers into the ongoing literature.

When building believable synthetic characters, it is key to their success that they behave in accordance with our theory of mind, and the intentional stance is the basis of that theory. Building synthetic characters requires a decision making architecture, and fortunately such things are well understood in computer science. The Belief, Desire and Intentions (BDI) architectures are explicitly based on folk psychology.

The BDI algorithms

The BDI algorithms explicitly implement folk psychological reasoning for rational agents [Bratman et al]. The Basis of the model is that a rational agent should reason from goals to action, but that is not all. The agent should commit of a course of action in the same way as we

```
forever do
   update beliefs using sensors
   update actions based on current plans
   if( a plan, P, has succeeded )
     mark relevant goal, G, as achieved
     drop the intention to achieve G
   elseif( a plan, P, has failed )
     if( there is another plan, P2, that might achieve G )
       replace P with P2 in the intention to achieve G
     else
       mark relevant goal, G, as failed
   endif
   if( there is a new goal, G )
     if( there is a plan P that might achieve G )
       form an intention to achieve G, using plan P
     else mark G as failed
 end //forever do
```

Figure 2: A BDI algorithm for a rational agent.

humans do. For an extended discussion of BDI as an algorithm see Wooldridge [2000], but a version of the algorithm is given in Figure 2.

The general description of the algorithm is as follows. First, given the agent has an explicit desire, such as "eat toast", the algorithm looks in a

plan library to see if it knows how to make toast. If there is a plan, P, then form a data structure called an intention, that has the goal "eat toast" and the plan P. Start executing the plan. If the plan fails - there is no bread on the table for example - then don't give up, see if there is another plan, P2, for satisfying the current goal. If so, then replace P with P2 in the existing intention. Things to note about this algorithm are that goals are explicit, and separate from plans, and that intentions form a commitment. The primary advantage of this algorithm over more naive approaches to rational agency is that the formation of commitment allows the agent to balance reactive (following a preset plan) with deliberative (thinking about another course of action) behaviour.

A secondary advantage is that BDI makes agents work the way our brains expect rational agents to work - it implements the intentional stance. Imagine a robot that wants to leave the house to post a letter, but finds that the the front door is locked. Many architectures (including BDI) allow for sub-plans and the robot might look for the key. If that fails however, a robot running BDI would then look for a new plan to leave the house - by using the back door perhaps. Other approaches might also find this solution, but without commitment, they might equally get distracted by another goal such as making toast or watering the plants. The key point is that we humans will expect the robot to still want to leave the room and if we observe behaviour other than that then we will think there is something wrong. Our wetware will start explaining away the observed behaviour of the agent - has the robot forgotten that that it wanted to leave? Does it not know that there is a back door? In fact the apparent problem is simply that it did not commit to that goal in the way we humans expect. It might still want to post the letter, but be quite happy watering the plants in the mean time. Such behaviour is rational, but it is not how we do it. More importantly, it is not how we humans expect rational agents to behave. It seems we humans have a hard-wired preference for interpreting an agents actions in terms of incorrect beliefs rather than rapidly changing goals.

Extending BDI

When Damasio argues that emotion is key to sensible behaviour, he is arguing against the rational behaviour of Mr Spock. The rational behaviour of folk psychology incorporates emotional responses to things and is the basis of, not just our predictive abilities of other people, but also of the behaviour of believable cartoon characters, the families of giant bugs, to dogs, cats, and robots. Damasio is right in saying there is more to it than rational action, but it is all built on this underlying level of common-sense [Norling]. To take an example from the entry in the

Stanford Encyclopaedia of Philosophy on folk psychology [Ravenscroft] "we remark that the smell of freshly baked bread made Sally feel hungry; that Sally wanted to go on a diet because she thought that she was overweight; and that Sally went to the fridge because she desired a piece of chocolate cake." Such reasoning introduces terms that refer to things that are not physical nor indeed detectable except in the context of reasoning about the behaviour of others. Sally "feels hungry", "wanted to go..", "thought that..." and "desired cake". Emotions and other mental attitudes are, by an externalist account of other minds, a means to explain behaviour within the framework of agents that do what they believe is in their interests.

A robot with mental attitudes

One can do an analysis of the motivations of Shakespearian characters, but these are complex and designed to exercise the mind of an adult. Instead lets look at the story-telling equivalent of "blocks world" and look at the children's Television show Teletubbies. The four Teletubbies live in a big dome located in a setting of rolling green hills, full of flowers and bunnies, with an occasional fluffy white cloud in a blue sky with a smiling sun. Their life consists of eating, sleeping, television, and the occasional sponge bath. That is it; the important thing is that they "love each other very much!" and do lots of hugging. The dialogue of the show consists of a voice over, one or two word utterances from the teletubies themselves, oh and lots of jumping up and down with excitement. In order to maintain this idilic existence, the Teletubbies have a favourite friend, the Noo Noo, which to quote the cover of the DVD [Teletubbies] "This friendly vacuum cleaner is very good at tidying up after the Teletubbies when they make a mess, but sometimes he just can't help being naughty. The Teletubbies love the Noo-noo, and Teletubbies love each other very much! Big Hug!" One Teletubby mess that keeps recurring is spilt tubby custard. A common scenario is that the Noo-noo will be doing some light dusting, sort of wandering about, and then see a blob of custard on the floor or wall. From the description of the Noo Noo, we, as observers, know that the Noo Noo wants to tidy up. The Noo Noo has a desire to tidy. When the Noo Noo ignores a custard blob, we know that the Noo Noo has not seen it. There is the suspense of waiting for the Noo Noo to notice the custard, and hence update its beliefs about the state of the world. When the Noo Noo does see it, there is some snout waving in anticipation of custard - the action is explained away by deciding that the Noo Noo has the mental atittude 'excited', and we know that the Noo Noo has the intention to tidy the spilt tubby custard. The custard is sucked up, and the narrative has closure. In this story, emotion is key, but it is an

explanatory mechanism on top of the common-sense reasoning of folk psychology.

In another story the tubby custard dispenser is broken (design stance) and won't stop producing custard. In this case the Noo Noo is confronted with a huge blob of custard, and does not pick it up. Why? We know the Noo Noo has seen the custard - we can see where it is looking, and everywhere it looks, there is tubby custard. So by our intentional model of the Noo Noo, it should pick it up. But the Noo Noo doesn't; Why not? This is not an academic question, it is a question raised and addressed by our wetware. The question pops into our head, and our brain automatically starts searching for an explanation. In this case watching the Noo Noo backing away, the answer is obvious - obvious to an agent with an anterior paracingulate cortex - the Noo Noo is scared. The tubby custard is too big, and the Noo Noo is worried.

Emotion plays a part in believable agents but it is within the framework of rational action. In the above description of the Noo Noo, we explain away the Noo Noo's actions in terms of beliefs, desires and intentions, and when that fails we introduce other mental attitudes to explain the observed behaviour. De Rossis has talked of BDI&E as an agent architecture for believable agents [Cavalluzzi] and it seems that approach is not just an engineering solution. The BDI&E mechanism would not only work the way we expect, but fail in a way we expect. To say that an agent is scared of big blobs of tubby custard, would have predictive powers about the agent's behaviour in future situations.

My Noo Noo

The Noo Noo in Teletubbies is a fictional character, but its mental life and physical capabilities are quite limited and hence amenable to being a real robot. My Noo Noo (Figure 3) is a real autonomous robot that cleans our kitchen. It uses an Intelibrain card [ridgesoft] to drive two electric drive motors, two servos, and the motor and fan from a Black&Decker Dustbuster. It uses BDI as the controller in a Behaviour Based Robotics architecture [Arkin] and uses a map of our kitchen to find its home and to remember where it will find the dirtiest spots [Wallis'06].

An observation from this real robot, situated among real children, is that the snout is recognisably a sensor. Children ascribe mental attitudes to the Noo Noo and it seems automatically map from the Noo Noo's form to that of more familiar rational agents. It is early stages but I hypothesize that importance of recognisable facial features on a believable agent is

significantly reduced when the agent has recognisable intentional behaviour.

But what about the qualia?

Humans are social animals and use folk psychology (with emotion) to predict the behaviour of others, and to help them deal with difficult neighbours. But this does not explain why we feel emotions ourselves. One can imagine a human zombie that acts like a real person but doesn't experience excitement, fear, and so on in the same way as we do. As a zombie, its behaviour is by definition indistinguishable from that of real humans, and thus there can be no evolutionary

Figure 3: My Noo Noo in our kitchen

advantage to being a real human. So why is the world not (presumably) full of zombies? [Chalmers]. The proposal - in need of fleshing out - is that we feel emotion as part of an attention mechanism. Seeing the Noo Noo back away from the very large blob of tubby custard, mirror neurons in our wetware make us do the same action. The last time we did that action, our attention was demanded by a large scary thing and we felt scared. Looking at the Noo Noo, I now have an emotion that will explain the Noo Noo's non rational behaviour of not picking up the custard when we know that the Noo Noo tidies up after the Teletubbies make a mess. The feeling of an emotion is there in order to bring events into conscious experience. Once there, our rational brain can reason about, not just the event, but also about other agents in the context of similar events. Such

reasoning is part of the way we get on with other people, and is thus key to the survival of the species. Folk psychology is a tool we use with other agents, the feeling of emotion is part of the mechanism that maps other's behaviour onto mental attitudes in our folk psychological model of the other's mental attitudes.

Conclusion

In conclusion, believable characters are those that act in accordance with our theory of (other) mind. The BDI algorithms provide a model of what that behaviour looks like, and more importantly, what it should not look like. An agent that too readily drops goals for instance will not behave as expected and will hence jolt us from our "willing suspension of disbelief". Emotions are certainly part of that theory, but the key is that other minds have a base level rationality, and mental attitudes such as scared, excited, tired and so on are introduced to explain away non rational behaviour. Of course human agents might not actually have beliefs desires and intentions to produce the observed behaviour, but that is how we think about them, and it is probably the easiest way to program that behaviour. Finally, in this paper I propose that we feel emotion in order to bring them to the surface where they can be reasoned about, and ascribed to others.

References

Arkin, Ronald C, Behavior-Based Robotics, MIT Press, 1998

Bates, J., The role of emotion in believable agents "Communications of the ACM", 37(7) pp122-125, 1994

Bratman, M. E., D. J. Israel and M. E. Pollack, Plans and resource-bound practical reasoning, Computational Intelligence 4, pages 349-355, 1988

Cavalluzzi (Addolorata) Valeria Carofiglio and Fiorella de Rosis, Affective Advice Giving Dialogs presented at "Affective Dialogue Systems", http://www.di.uniba.it/intint/people/papers/ADS'04.pdf, 2004

Chalmers (David) The Conscious Mind: In search of a fundamental theory, Oxford University Press, 1996

Damasio, Antonio R., Descartes' Error: Emotion, Reason, and the Human Brain, Avon Books, 1994

Dennett, Daniel C. The Intentional Stance, MIT Press, 1987

Ephron, Nora (Director), Sleepless in Seattle, 1993

Gallagher, H., A. I. Jack, A. Roepstorff and C. D. Frith, Imaging the intentional stance in a competitive game, Neuroimage 16, pages 814-821, 2002

Humaine NoE http://emotion-research.net/aboutHUMAINE/, July 2006

Imai, M. T. Ono and H. Ishiguro, Robovie: Communication Technologies for a Social Robot, "International Journal of Artificial Life and Robotics", 6, pp73-77, 2003

Meadows, Mark Stephen, Pause & Effect, the art of interactive narrative, New Riders, 2003

Norling (Emma) and Liz Sonenberg, Creating interactive characters with BDI agents, Proceedings of the Australian Workshop on Interactive Entertainment IE2004, Yusuf Pisan (ed), Sydney, Australia, Feb 2004

Ravenscroft, Ian, http://plato.stanford.edu/entries/folkpsych-theory/, 2006

Ridgesoft, http://www.ridgesoft.com/

Teletubbies Naughty Noo-noo! (DVD) CBeebies BBC, 2004

Wallis, P. Helen Mitchard, Damian O'Dea and Jyotsna Das, Dialogue Modelling for a Conversational Agent, AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence", Markus Stumptner, Dan Corbett and Mike Brooks, 2001, Springer (LNAI 2256)

Wallis, P. The Symbol Grounding Problem for Two Year Olds, International Conference on Computers and Philosophy (i-CaP), Laval, 2006

Wooldridge, Michael Reasoning about Rational Agents, MIT Press, 2000

AWARENESS, ACTION AND ATTENTION

JACQUES PENDERS

Research Group on Mathematical Linguistics

Rovira i Virgili University

Tarragona, Spain[157]

Abstract

One often assumes that we, human beings are rational and first think and than act. This paper is an attempt to describe the mental characteristics governing the performance of regular everyday actions; and shows that no mental act has to precede our actions, instead of consciously thinking before we act, we mostly act while simultaneously overseeing our acting. The case of ball juggling is used to underpin the analysis with empirical facts.

Introduction

In this paper I make an attempt to describe the mental stance applied by a human being while performing the standard routines of everyday life. The drive behind this attempt stems from research into what is called

157 Currently Sheffield Hallam University, Sheffield UK. +44 114 225 3738. j.penders@shu.ac.uk

Machine Consciousness studies. Machine Consciousness studies cover efforts to construct machines that display characteristics which one might call mental.

The notions of mind, the mental and consciousness have been studied extensively in Philosophy. For the purpose of sketching the position of my research a brief discussion of some of the central notions. Concerning the duality of body and mind the dominant presupposition of western thinking is that body and mind are distinct and that the mental realm is distinct and separate from the material world in which the body acts. Descartes' famous '*I think therefore I am*', is often taken as to imply a notion of self-consciousness. Self-consciousness is often thought to be manifested as rationality: to be able to reason about oneself. Moreover, rationality is usually associated with the verbal, resulting in associating mental processes with a language of thought. The presupposition that mental processes proceed as a language of thought tempted some philosophers to define consciousness as a 'Centre of Narrative Gravity', (Dennett 2002).

In line with such reasoning is the often-encountered assumption – which I believe is generally untrue – that a certain mental act precedes our bodily actions, or in plain language that we first think and then act. For instance Haggard et al. (2002) write: "*Normal human experience consists of a coherent stream of sensorimotor events, in which we formulate intentions to act and then move our bodies to produce a desired effect*".

Indeed, on occasions we do first think and then try to act accordingly. Being human, we like to think of ourselves as rational beings. In the history of Philosophy Immanuel Kant is probably the clearest exponent of this view. He saw a human being as a logical subject of thought (Stuart, 2005) that is bound to act in the physical world. Kant's work could be seen as a major attempt to give primacy to rationality. However, Kant did not assume that we are rational; he argued that we should be rational; in his view rationality had to be fought for.

Concerning mental processes and body control, William James (1890) clearly noted that the suggested ordering in time in which a mental act precedes our bodily actions, does not hold. He described his concept of ideomotor action summarised as: we think the act and it is done. An example of his: "*We think to drink our coffee and we find ourselves already holding the cup in our hands*". I will argue a step beyond and show that we often act before any conscious thinking can occur. My point is not to substantiate a general moral excuse for cases where we have done things, which we afterwards regret. My point is pragmatic: we cannot act and behave as we do in ordinary life if we first have to think (let alone think over) every action.

Whereas philosophy of mind generally analyses and then tries to explain the working of the mind, machine consciousness studies aim at a constructive approach. Machine consciousness studies are considered a branch of robotics or Artificial Intelligence. Chrisley et al., (2005) describe the aim of these studies as: 1) to create artifacts that have mental characteristics typically associated with consciousness (such as awareness, self-awareness, emotion and affect, experience, phenomenal states, imagination etc.) and 2) to model these aspects of natural systems in embodied models (e.g., robots).

This definition stipulates that the mental phenomena are to be studied in an embodied creature; thus the combination of computing machinery for information processing and mechanical actuators generating physical action is brought into the focus. My aim is neither to discuss whether the aims of machine consciousness studies can ever be achieved using the means currently applied in machines and robots; nor whether are the aims of Artificial Intelligence and robotics achievable. My interest is in whether these constructive approaches produce new insights. Present-day computers provide robots with tremendous reasoning capabilities. Nevertheless, the currently applied robots are far from being able to perform actions which seem elementary to a human being, such as throwing and catching a ball. Robotics and Artificial Intelligence move into an area, which is so familiar to us that we assume it all to be obvious. The lesson to be learned is that we hardly understand how we ourselves perform our actions, and in particular how sensory inputs guide our actions.

The present paper is an attempt to identify the mental processes, which manifest themselves in regular action oriented contexts. Without being able to provide a systematic view, I will discuss a few assumptions which indicate and position the relevant mental processes. Reasoning appears not to provide the solution for building artificial creatures and I will show that rationality is not the major guide for our everyday acting; obviously, the latter does not imply that irrationality applies. Rationality requires reasoning and reasoning is a conscious process; below I show that conscious processes cannot control our everyday actions.

In this paper I will hardly touch upon the notion of consciousness. Instead, the line of reasoning is the following. Any action and generally any perceptual input is accessible to consciousness only if it has passed through or has been passed on by attention. The processing by attention takes time and causes delay, and if the control of all actions has to pass attention then certain performances cannot be executed by a human being. Nevertheless humans do perform these. Thus, the control of our acting does not necessarily pass attention, which means that certain activities and performances are beyond the control of consciousness. Instead of first thinking and then acting, we often only oversee our actions

with our conscious and rational minds. Below, I will use juggling as the primary example to investigate the flow of the mental processes and prove my point.

Attentional focus and Acting

In the morning of a regular day, while deliberating on how to make the best out of the day ahead, we routinely drink our coffee and make our way to work, say by car. While driving the car, we suddenly stand on the brakes as we are forced to an emergency stop. Only after having come to a standstill it becomes clear what has happened the seconds before and what has been our contribution to the event.

My interest is in the mental stance governing the behaviour before the emergency stop, and which I believe we usually take when routinely drinking coffee or driving the car. This is a stance under which actions are selected and performed (for instance grabbing the coffee cup or pushing the brakes) without the actions being in the focus of attention.

In order to explore the stance, first a few words about the notion of attention. Our mind can be in different modes of activity, with sleeping as the extreme on the less active end. When awakening from sleep, our mind has to "warm-up" in an arousal phase. Then we become generally aware enough so that we can attend: the mind is aroused and proceeds via getting aware to attention. Further onwards, when there is attention, conscious experiencing, and consciousness and reasoning may come in.

Our senses produce an overload of signals as they are continuously subject to various stimuli. Broadbent (1958) argued that the processing of semantic features (´features related to the meanings of objects´) from the senses' inputs has severe capacity limitations. And since Broadbent's work the faculty of attention is often conceived of as a filter for or a gate to consciousness, which blocks, weakens or inhibits incoming messages from the senses. Baars (1997) introduced the metaphor of attention acting as a spotlight in a theatre. When in the spotlight of attention, the mental processing becomes accessible to consciousness. The filter metaphor characterises the operations of attention as reductive while the spotlight metaphor suggests amplification; both nevertheless agree that attention is the gateway to consciousness and that it is selective.

A different but equally important aspect of attention is that it also has to do with action. "Awareness [or being aware] implies perception, a purely sensate phase of receptivity. Attention reaches. It is awareness stretched toward something. It has executive, motoric implications. We attend **to** things." (Austin, 1998).

Attention can guide our actions; however, the question is whether all our actions are guided and controlled by attention. Appropriate applications of motor skills - that is to act appropriately - requires a proper combination of perception, action selection and action execution. The role

of attention in relation to perception has been widely studied; however its role in applying motor-skills has not received as much scientific interest. The reason for this might be that motor-control, which is a prerequisite for motor-skilfulness, is very much on and below the edge of what we can consciously experience and control.

The performing arts and sports sciences deal with action and attention. Artists and sports men and women engage in what is called *deliberate practice* (Rossano, 2003) (Ericsson et al., 1993): the concentrated effort to hone and improve specific (mental and) physical skills. Literature on deliberate practice distinguishes between external attentional focus and internal attentional focus; internal attentional focus means that the performer directs attention to the movements itself, while in external attentional focus, the attention goes to the effects the movements have on the environment (Wulf and Prinz, 2001). Obviously, in both attitudes attention plays a prominent role.

The influence of internal attentional focus may be observed in, for instance, dancing or martial arts classes. In a class of beginners, the students might be quite able to follow and copy the movements of their instructors. However, when the instructor explains the consecutive moves to the very detail, several students appear not to be able to perform, even though they may have performed quite well before. The reverse also applies: when the instructor is asked about the details of a move which (s)he has never made explicit before, it is likely he or she has to perform the movement first before being able to explain. Applying internal attentional and conscious focus to motor-control hampers the performance. Extreme examples are observed with patients suffering from the syndrome called apraxia. Apraxia denotes the inability of a patient to perform a certain skilled movement. For instance when asked to demonstrate teeth brushing, the patient is unable to do so, whereas he or she is perfectly able to brush the teeth in the morning, when there is no particular emphasis on the act itself.

Attention obviously has motoric implications. Generally internal attentional focus slows down movements and external attentional focus is more proficient (Wulf and Prinz, 2001). The examples show that internal attentional focus and conscious control of motor-skills may even lead to an inability to act.

The notion of external attentional focus is not clearly defined and allows several interpretations. In a narrow, but easy to define sense it denotes attention focusing on bringing about a single effect: directing a tennis ball, or throwing a single ball or bean bag into the air such that it can be caught. I will test this reading in the context of juggling.

When performed well, juggling is great to watch. The magic about it is that the general spectator perceives the pattern that is formed by the balls (or other objects), but neglects the movements of the juggler. For instance, in the three-ball pattern called yoyo (Dancey, 1994) the juggler

only throws two balls while the third ball is kept in one hand and is carried throughout the pattern, but that does not in the least bothers the spectator. Only when the spectator is observing the scene as a whole but is neither focussing on a particular ball nor on the moves of the juggler the typical yoyo effect –of the balls appearing to be connected- is perceived. However, if one focuses on the one ball that the juggler keeps in the hand, the yoyo effect disappears.

In the literature on sports psychology it is often assumed that external attentional focus is the only alternative to internal attentional focus; refer for instance to (Wulf and Prinz, 2001). The stance of the spectator shows that a third stance, one without focus, is possible as well. In the next section I will first investigate whether external attentional focus applies for a juggler; as the answer will be negative, I will also explore whether a stance similar to that of the spectator might apply.

Acting and Attention Shifts

The basic pattern in five-ball juggling is the cascade; and although basic, it is quite a step beyond three-ball juggling (Dancey, 1994). It is hard to learn and requires fast acting, the complication being that between throwing and catching the same ball four other objects – three of which are already up in the air - have to be handled. When first starting, it is a problem to throw each of the five balls one after the other before the first has returned ('flashing' as it is called), in doing so a novice will not be able to tell which ball was first thrown, let alone be able to catch it with the proper hand. The novice juggler is trying to apply full and focussed attention, and that leads him or her astray.

Juggling requires fast series of combinations of perception, action adaptation and action. The handling of a single ball is cyclic. To estimate the time of a cycle, assume a throwing height of one meter, which is more then most patterns, including the five-ball cascade require. The law of gravity leaves less than a second of time between throwing and catching the same ball. In the interval of less than a second that this one ball is going up and down, four other objects have to be dealt with. They are flying around and have to be observed in order to be handled; figure 1 gives an overview of the five-ball cascade.

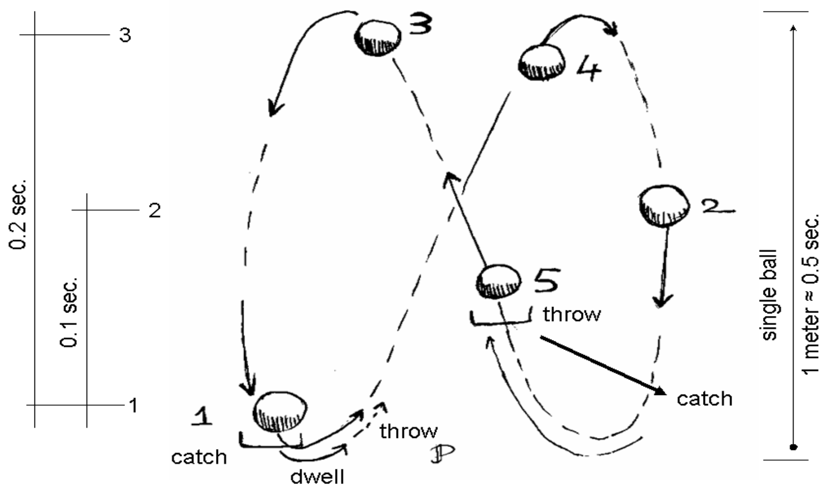


Figure 1, the main actions in a five-ball cascade.

Observations of jugglers show that the time lapse between two catches of the same hand may be as little as 0.2 seconds (refer to Polster (2003) for more details), this is indicated in figure 1 for consecutively catching balls (1) and (3). In this short interval several actions of this hand merge into each other: catching, bringing to throwing position (dwelling), throwing and preparing/waiting for the next. Since the juggling pattern is regular: in the middle of this series the other hand has to start its own series as well. In figure 1 ball (1) was the last to be caught and ball (2) is the next coming down and is to be caught about 0.1 second after ball (1) was caught. Obviously, the exact time between throwing ball (1) and catching ball (2) is less then 0.1 second; unfortunately I have no exact data.

In order for attention to be in control of the juggling actions, attention has to shift from ball to ball. Broadbent (1958) concluded that per second no more then two attention shifts can occur. Currently, psychologists distinguish between voluntary or internally (endogenously) driven attention shifts and involuntary or externally driven shifts; Broadbent only considered voluntary attention shifts (Lachter et al. 2004). Nevertheless, if attention is in control, the shifts have to be voluntarily. Recent work has found variations for voluntary attention shifts from 0.5 second to 0.15 seconds (Lachter et al. 2004). Obviously, even a fast voluntary attention shift of 0.15 second is too slow to switch from ball (1) to ball (2) since the latter is due within 0.1 second. Voluntary shifts would suffice to handle the balls due for one hand, but are too slow to interweave the actions of the second hand.

The previous section showed that internal attentional focus hampers execution of actions, and that the resulting actions are generally slower than when external attentional focus is applied. The case of five-ball juggling shows that external attentional focus, in the sense of attention focusing on bringing about a single effect fails as voluntary attention shifts require more time then the pattern allows.

Attention Reduced

Five-ball juggling cannot be governed by focused attention and voluntary attention shifts. Alternatively, involuntary attention shifts turn out to be much faster. Involuntary shifts require only 0.05 second (Lachter et al. 2004), which is in the order of three to ten times faster. The research into attention shifts is mainly based on experiments in which subjects are exposed to visual or audio inputs and are asked about what they perceive. However, juggling consists of an intricate combination of perception with actions such as movements of the limbs. Concerning acting, one also distinguishes between voluntary and involuntary acts. Similarly to the differences for attention shifts, voluntary acts are also slow compared to involuntary acts. The time required for the single voluntary act of pressing a button *only* when a light flashes is about 0.15 seconds (Austin, 1998). Voluntary acts are too slow to meet the constraints of juggling, thus, juggling is neither a series of voluntary actions. Involuntary acts, for example a reflexive jerk to shield the eyes from a flashing light, take only 0.025-0.05 seconds (Austin, 1998). Note that the times required for attention shifts are in the same order of magnitude as the times required for acting. The latter suggests that on occasions the body is as fast or maybe even faster than the mind.

Juggling is of course not a series of reflective jerks; the point is that humans can execute perception–action cycles at high speed. A recent assumption in cognitive neuroscience is that the mind has a layered structure with different organising levels concerning body experience. Neuroscience has found that there exist several distinct neural systems or circuitries for the perceptual control of movement (Rossano, 2003) and (Waszak et al., 2005). Raichle (1997) makes a distinction between “the neural circuitry underlying the unpractised, presumably conscious performance of a task on the one hand, and the practised presumably nonconscious performance of a task on the other hand.” The response time of the latter circuitry is significantly shorter than that of the first (Raichle, 1997). More recently, Waszak et al. (2005) distinguish between actions carried out in response to exogenous stimuli or *stimulus-based* actions, and actions selected endogenously or *intention-based* actions. They note that intention-based actions are typically goal-directed, but slower than stimulus based actions.

In order to meet the time constraints, a five-ball juggler must apply a mental stance differing from internal or external attentional focus. This stance avoids intention-based actions, allowing the neural circuitry for stimulus-based actions to perform. Nevertheless, the stance must be very sensate and requires awareness; lacking an appropriate name, I call this stance: **non-focussed awareness**. And indeed, the experienced juggler does not focus on the individual balls. In his juggling book Dancey (1994) advises: "*While learning [a five-ball pattern] you are trying to* make *yourself do it, when you can do it you* watch *yourself doing it."* When acting, the juggler seems to be in a stance, which to a certain extend resembles that of a spectator.

As said before, to learn to juggle five-balls is hard; the above observations help to explain this fact. The time constraints are too tight to apply conscious controlled or voluntary actions, nevertheless it is the slower intention-based circuit that is applied to learn or correct a move.

I have shown that there simply is insufficient time for attention to interfere in five-ball juggling and that restricting attention results in faster actions. The surprising thing is that when no full attention is required for acting, the mind performs other tasks concurrently.

Three-ball juggling is less demanding than five-ball juggling. While juggling, the juggler can do other things as well, for instance speak, walk etc.; however non-focussed awareness is permanently required, when the juggler's attention drifts away and focuses elsewhere the balls drop. Car driving implies a similar requirement; the driver can perform many other things while driving but a certain level of awareness is required throughout. In daily life we perform many actions without attentional focus, car driving and juggling are two of the many possible examples, cycling and walking are others. For instance, when walking the body performs an intricate combination of muscle activities to maintain posture.

I have avoided any attempt to define the notion of attention; therefore I cannot conclude that attention is not involved in the stance of non-focussed awareness. But referring to the spotlight metaphor, if there is attention involved, it is only a dim light. Because attention is a preliminary for consciousness this conclusion has implications for consciousness as well.

Concerning the relationship between consciousness and the body, the notion of body image plays a central role. In Yamadori's three-layer model of the mind the body image emerges at the highest level. "The lowest level is an assembly of neuronal information coming from all parts of the body; at the middle level the body schema are situated which secure the emergence of the conscious body image at the third level" (Yamadori, 1997). The body schemata are subsystems 'implementing' James' ideomotor actions, for instance grabbing the coffee cup. The

suggestion is that the body image generates at the middle level and may pass on to the conscious level, thus leaving no active role for consciousness. The second level is rather independent from the conscious third level, which is confirmed by the split-brain studies and in particular very compellingly by the so-called *Anarchic hand* (Blakemore et al., 2002). The latter designates pathological behaviour in which a patient's right hand manipulates a tool properly but 'spontaneously', that is without the patient neither consciously initiating the movement nor being able to inhibit the action. The anarchic hand shows that neither attention nor consciousness is a prerequisite or a necessary condition (sine qua non) for action; neither if them is necessarily the initiator of actions. Moreover, it even shows that there exist pathological cases where consciousness is unable to inhibit actions.

Most people readily acknowledge that the internal functioning of our body is beyond our conscious control. The anarchic hand shows that even skilful behaviour might be beyond the span of control of consciousness.

Conclusions

Using the case of five-ball juggling I have made an attempt to analyse the mental stance taken by a human being when performing. I have called this a stance of *non-focussed awareness.* Unravelling this stance is interesting on its own, but it also sheds some light on the complex of mental states and stances by which a human being monitors and controls his or her body and actions. Definitely the human body on its own is a complex system with a complex control structure, the understanding of which could function as a paradigm for robot and machine design.

The case of five-ball juggling showed that the often-supposed sequence that a mental act precedes bodily actions - or that we first think and then act - cannot hold. Juggling is not a series of voluntary actions governed consciously.

The conscious processing capacity is limited; attention is a gate to or filter for consciousness. Acting requires perception, action selection and action execution. Shifting the focus of attention is a relatively slow process. Five-ball juggling would be impossible if focused attention has to be applied. In general, attention may interfere with acting but that often results in poorer or slower execution. Restricting attention results in faster actions. The mental stance of the juggler is a very sensate stance in which there is typically little or no attentional focus. The latter implies that the performance cannot be subject to conscious control.

Besides juggling, many actions are as well initiated and performed without attentional focus; they are mostly on and below the edge of conscious experience and control. In everyday practice we usually act before consciously thinking of it. The occasions where thinking precedes

acting are the exception and not routine practice. Consciousness not necessarily initiates actions, moreover in certain cases conscious control cannot inhibit actions. Consciousness only has weak control over the acting body, even though subjects have the feeling they consciously control their body. Nevertheless, we do oversee our actions with our conscious and rational minds and except for pathological cases we are able to suppress many 'spontaneous' actions.

References

J.H. Austin, *Zen and the Brain*, MIT Press 1998.

B.J. Baars, In the Theatre of Consciousness; The Workspace of the Mind, Oxford University Press 1997.

S-J Blakemore, D.M. Wolpert and C.D. Frith, Abnormalities in the awareness of action, *TRENDS in Cognitive Sciences* Vol 6, no 6, 2002.

D.E. Broadbent, *Perception and Communication* New York: Oxford University Press.

Chrisley, R., Clowes, R. W., & Torrance, S. "Next-generation approaches to machine consciousness". In R. Chrisley, R. W. Clowes & S. Torrance (eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment*, 2005.

C. Dancey, *Encyclopaedia of Ball Juggling*, Butterfingers, Bath UK 1994.

Ericsson, K.A., R.Th. Krampe and C. Tesch-Römer, 1993, 'The role of deliberate practice in the acquisition of expert performance.' *Psychological Review*, 100: 363-406.

Patrick Haggard, Sam Clark and Jeri Kalogeras. Voluntary action and conscious Awareness. *Nature Neuroscience* volume 5 no 4. 2002

W. James *The principles of Psychology*, 1890; Harvard University Press 1983.

Lachter, J. K.I. Forster and E. Ruthruff, Forty-Five Years After Broadbent (1958): Still no Identification Without Attention, in *Psychological Review* Vol. 111, No 4, pp880-913, 2004.

B. Polster, *The mathematics of Juggling*, Springer-Verlag 2003.

M.E. Raichle, Automaticity: from reflective to reflexive information processing in the human brain, in: *Cognition, Computation and Consciousness*, K.Ito, Y. Miyashita and E. Rolls (eds), Oxford University Press, 1997.

M.J. Rossano, Expertise and the evolution of consciousness, *Cognition* Vol 89, (3) 2003

Singer, R. N. (1985). Sport performance: A five-step mental approach. *Journal of Physical Education & Recreation*, **57**, 82-84.

Singer, R. N. (1988). Strategies and metastrategies in learning and performing self-paced athletic skills. *Sport Psychologist*, **2**, 49-68.

Susan Stuart, The Binding Problem: Induction, Integration and Imagination, ". In R. Chrisley, R. W. Clowes & S. Torrance (eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment*, 2005.

F. Waszak, E. Wascher, P. Keller, I. Koch, G Aschersleben, D. A. Rosenbaum and W. Prinz, Intention-based and stimulus-based mechanisms in action electionIntention-based and stimulus-based mechanisms in action selectiong. Experimental Brain Research Volume 162, Number 3 / April, 2005.

Gabriele Wulf and Wofgang Prinz Directing attention to movement effects enhances learning: A review, *Psychonomic Bulletin & Review*, Volume 8, Number 4, 1 December 2001, pp. 648-660(13)

A. Yamadori, Body awareness and its disorders, in: *Cognition, Computation and Consciousness*, K. Ito, Y. Miyashita and E. Rolls (eds), Oxford University Press, 1997.

ARTIFICIAL AGENTS AND THEIR ONTOLOGICAL STATUS

Denis BERTHIER

Professor (Artificial Intelligence and Epistemology)
Institut National des Télécommunications
Groupe des Ecoles des Télécommunications
9 rue Charles Fourier, 91011 Evry Cedex, France
phone: 33 1 60 76 41 22
mailto:Denis.Berthier@int-evry.fr
permanent URL: http://www.carva.org/denis.berthier

Abstract : Can an artificial agent "really" think? can it be "really" intelligent ? can it "really" have beliefs, goals, intentionality? Such questions have been plaguing AI since its inception. The solution we propose is to assert that an "artificial agent" is a *virtual* agent – and that all the cognitive or mentalistic attributes we may be tempted to grant it are also virtual. Whereas the first qualifier is purely descriptive, the second one is about the ontological status of such things as appear to us as agents. Our solution relies on: 1°) eliciting a precise core meaning for the word "virtual", a meaning that differs radically from the one used by philosophers (Bergson, Deleuze,…) but that has always been implicit in specialized contexts ("virtual image", "virtual world") and that should now be given its full ontological generality; 2°) relating the virtual to a broadened notion of interoperability, which justifies our assertion on technical and psychological instead of philosophical grounds. We relate this interpretation of AI to strong AI, to weak AI and to Dennett's intentional stance. We stress its implications for the cognitive sciences project of "naturalizing" intentionality. Finally, we mention some consequences in the information systems domain, relative to the acceptance of agent concepts for the modelling of organizations and their business processes.

Key Words : Epistemology, intentionality, Artificial Intelligence, virtual, interoperability.

Introduction to the problem and to the proposed solution

A question has been plaguing for decades all debates about Artificial Intelligence (AI), sometimes in so heated and inconclusive ways that it seems nearly everybody is sick with them and nothing new is to be expected. Can an "artificial agent" be "really" intelligent? can it "really" think? can it "really" have knowledge, beliefs, goals, intentions, emotions and so on? can it "really" have intentionality (in the full phenomenological, Husserlian sense), i.e. can it "really" refer to things in the world out there *and* to things in our heads? can it "really" communicate with us, share knowledge with us, collaborate with us – and what does all this mean?

Disregarding all the hype about AI and all the frustration it generated, the main dilemma initially raised by such questions is not only still present; it gets stronger than ever as AI goes deeper and deeper into the simulation of complex human behaviours:

– on the one hand, two very strong and complementary arguments, one technical and one psychological, plead in favor of an agent oriented vision : a) in order to design some complex software systems, it is very helpful (or even necessary) to consider them as agents and to use some related formal mentalistic concepts[158]; b) moreover, in the proper context (i.e. in the operation framework it has been designed for), such an "artificial agent" is spontaneously perceived as having the aforementionned mentalistic attributes[159]; as a consequence one cannot be satisfied by just blaming AI (as is often the case) for a lack of precaution in its vocabulary;

– on the other hand, for philosophical, religious or common sense reasons, very few people are willing to "really" grant mentalistic attributes to a machine, however sophisticated it may be.

158 As to the practical scope of this remark, let us remind that, giving its full meaning to (Newell, 1982), sophisticated methodologies have been developed and are widely used in order to facilitate software development in terms of agents, such as Gaïa (Wooldridge & al., 2000) and KADS (Schreiber & al., 1993) or its multiagents version MAS-Common KADS (Iglesias & al., 1997).

159 "Every human being is so much predisposed by naïve psychology to conceive his actions and those of other people as the result of their goals, intentions, desires and beliefs that the least non human behaviour is irresistibly understood as that of an agent equipped with an intention or a goal" (Jacob, 2004, p. 13 – translation and italics are ours).

The solution we propose is to assert that an artificial agent is a *virtual* agent – and that all its cognitive or mentalistic attributes are also virtual. Here, the two qualifiers have completely different bearings: whereas the first one is purely definitional or descriptive, the second one is about its ontological status (where we understand "ontology" as being devoid of any form of essentialism). Moreover, if we decompose the expression, the adjective "artificial" means "produced by a technical activity", but the word "agent" does not suppose any specific definition of agenthood. So that *our assertion should be understood with the broadest scope: as soon as, in its limited operation context, a software component appears to us as if it was an agent (whatever our notion of agenthood may be, whatever the implementation techniques it relies on may be) and it ceases to do so outside this context, then it is a virtual agent; the same applies to any of the mentalistic attributes we may grant it*.

Our solution amounts to replacing a yes-or-no question by a question relating to ontological modalities in general. Admittedly, it is a nonsensical or a quasi void assertion and it is therefore a delusory solution, if one relies either on the standard meanings of the qualifier "virtual" (as they are recorded in the dictionnaries or as the word is used by most philosophers), or on its vague or nearly undefined meaning (as it is currently used on every occasion). Saying the agent is virtual may even not be completely new in this bare form[160]. However, what follows is new.

In the subsequent three sections, our assertion is first explained at three different levels of understanding: based on analogies, based on a new core definition of "virtual", based on the notion of interoperability. Then consequences are drawn regarding other interpretations of AI (strong AI, weak AI, Dennett's intentional stance) and the cognitive sciences project of naturalizing intentionality. The last section draws a consequence for information systems.

160 For instance, the FIPA (Foundation for Intelligent Physical Systems, the international association for the normalization of multi-agent systems) specifies in its norms that an agent has a virtual knowledge base; in this context, the word "virtual" may be understood with the meaning we have elicited, but FIPA has not noticed that such elicitation was necessary. We are not aware of any explicit claim about artificial agents as being virtual in any precise sense.

Level one: analogies

The example of a reflection in a mirror (which is technically a virtual image) is prototypical of the meaning we want to associate with "virtual". More generally, the following table (some terms of which will be explained later) lists the analogies we want to establish between the idea of an agent being virtual and the way this word is used in association with three other different phenomena. For each phenomenon, we display a short explanation and the framework defining its condition of possibility.

Phenomenon	Explanation	Operation framework
Virtual image	Propagation of light rays	Being in the proper light cone and limiting oneself to the visual modality
Virtual sound	Propagation of sound "rays"	Being in the proper spatial domain and limiting oneself to the audio modality
Virtual world	Sensorimotor interoperability	Being connected to the VR apparatus and limiting oneself to the predefined modalities it supports
Virtual agent	Semiotico-cognitive interoperability	Being in the proper predefined communication situation (language, topic, comm. links)

There are four major points these analogies intend to stress.

First, a virtual image or a virtual sound or a virtual world is not a real image or a real sound or a real world, but it *is* in all cases plainly *actual* – as opposed to potential. The only restriction is that there are conditions for my effective perception of it (to see a reflection in a mirror, I must be in the proper light cone and look in the proper direction), but it is a fact that the satisfaction of these restrictions is extrinsic to the situation itself, and therefore not of a kind such that one could find anything potential in this situation. Given a situation A of an object in front of a mirror, my

entering a situation where I perceive the reflection is an event that doesn't change anything about the mirror, the object or its reflection – neither in the way the event of a dam breaking could change the state of the water it held (and the flood status of the valley below it), nor in the way the event of a measure on a photon could change (i.e., in this case, determine) its polarization state. This event, which is undoubtedly relative to situation A, could be called possible or contingent on it, but, being totally external to situation A itself, it cannot be called potential (nor latent); nor can anything in situation A itself.

Second, in all of these examples, the notion of a "context of validity" or "operation framework" is inseparable from our notion of the virtual. Concerning an artificial agent, this must be related to the fact that it appears to us as an agent only in the proper communication situation; we are aware of no artificial system that had no severe limitations on its operating context – hype notwithstanding. Of course, there is much research effort to alleviate such limitations, but enlarged context does not mean unlimited context.

Third, in any of these examples, there is nothing subjective or imaginary (in the sense "folle du logis" often associated to the word "imagination").

Fourth, one could object that there is no predefined absolute reality and that anything could therefore be said virtual. But this would be absurd in all the previous examples, a virtual "object" being defined in opposition to its real counterpart. And there is no virtual world without a real aparatus in a real world to support it. We cannot say for certain what "real" means, but anyway virtual can be defined only in opposition to it.

Level two: eliciting a new core meaning for "virtual"[161]

Considering etymology and the previous first three examples taken from science and technology, one can elicit *a new core meaning for*

161 Our work on the virtual was inspired by Wittgenstein's claim that "philosophy aims at the logical clarification of thoughts" (in his "Tractacus") and by its complementary idea that, in order to dispel our confusions, we should inquire how our "language games" are used (in his "Philosophical Investigations" §115): "A 'picture' held us captive. And we could not get outside it, for it lay in our language and language seemed to repeat it to us inexorably." Here, the "picture" was the idea that a simple yes-or-no answer was needed to questions raised about AI (intelligent or not? …).

"virtual": that which is not real but displays the full qualities of the real, in a plainly actual – i.e. not potential – way. Contrary to the current meaning, *this definition distinguishes clearly the virtual from the potential* and *it allows to understand why the virtual can have real effects*.

The current standard meaning of "virtual"[162], inherited from medieval Scolastics (and from the invention it made of the pseudo Latin *virtualis*), entails "not in actual fact" and can therefore hardly be distinguished from "potential"; the word, with this meaning, has been used intensively by the French philosopher Bergson (in Bergson 1896/1970, but also in many other of his works); a kind of a theory of the virtual as a process was developed by another French philosopher, Gilles Deleuze, in "Difference and Repetition" (1968), claiming to formalize Bergson's notion (a claim that may certainly be debated); more recently, Gilles-Gaston Granger (1995) still adopts a similar meaning. Another current meaning of this word in colloquial usage is "nearly" or "quasi" or even "pseudo"; we consider it as mainly rhetorical, following a fashion associated with the information technologies boom.[163]

But both of these meanings are in total contradiction with the intended meaning in our previous examples, in expressions such as "virtual image" (in geometrical optics) or "virtual sound" (in the music or movies industries) or "virtual world" (in the "virtual reality" domain – VR). Although these expressions are recorded in dictionnaries, they have not yet been related to any general core meaning for virtual (and therefore they have no compelling philosophical implications).

On the basis of etymology, virtual is "what has the virtue of"; and virtue (from the latin *virtus*, itself derived from *vir* – man, hero) means quality with underlying strength. Therefore, following (Berthier 2004, 2005a, 2005c), we define "virtual" as "that which is not real but displays the full qualities of the real, in a plainly actual – i.e. not potential – way". It is then easy to check that the above examples satisfy this definition.

There is a major consequence: something virtual can have actual effects – for instance, one can cure agoraphoby by walking in the "open air" in a virtual world. Even if the world we are in at some moment is virtual, all the mental experiences and feelings we live in it are still fully

162 Trésor de la Langue Française, Encyclopædia Universalis, Wordnet…

163 One should also mention a regional meaning, in the information technologies domain: "digital" – which seems to us unnecessarily restrictive.

real. In ours analogies, this can be compared to the fact that the ligth rays "issued from" a virtual image are real. Any definition of the virtual that does not distinguish it clearly from the potential makes it logically very difficult (even impossible) to explain this fact: how could something that is potential and remains unchanged have actual effects? This would logically amount to ask how the water in a dam, which has the potential to flood the valley below, could flood it without changing its own situation.

Let us now consider our assertion that an artificial agent is a virtual agent. This is the place to notice that, at a first level, it rested on a vague analogy: the artificial system appears to us as if it was an agent, in the way a virtual image or a virtual sound appears to us as if it was real; but there are contexts were these appearances vanish; therefore let us call them virtual. What the present definition of the virtual provides is a second level of understanding of the assertion, i.e.: a) a meaning more precise than this mere analogy, because it involves instead a general ontological modality, defined from a general phenomenological standpoint; *and* b) a first phenomenological justification based on this meaning: it is a virtual agent because one can observe that, in its operation framework, it displays the full qualities of an agent.

Level three: justifying our assertion on the basis of interoperability

The next step provides a stronger justification for our assertion by explaining *why* the artificial agent appears to us as an agent – much as explaining (in terms of light rays) why a virtual image appears to us as a real image justifies calling it a virtual image. This is done by introducing a broadened notion of interoperabilty (which generalizes the technical meanings of this word in the information technologies domain) and eliciting its duality with our definition of the virtual.

This supposes an approach of AI that relies on its effective practices and results (Berthier, 2002, 2004) instead of on general claims of realizations to come "in the near future"[164]. And the effective results – the

164 An example of such hype is Lenat's announcement of e-Cyc and its capacity for automatic knowledge acquisition from the Web (Austin Chronicle, Dec. 19, 1999). Under the title "Cyc Invades Cyberspace", he writes: "When e-Cyc becomes fully operational in early January, a thermonuclear explosion in the amount of information being pumped into Cyc's knowledge base is expected, with the result of Cyc becoming exponentially

innumerable products that have reached industrial and commercial stage – are specialized agents that are designed according to precise methodologies to solve pre-specified types of problems in pre-specified operation frameworks (a fundamental notion, separating AI from science fiction[165]).

In this conception, *AI can be understood as aiming at developing semiotico-cognitive interoperability between Man and the computer (and VR as aiming at developing sensorimotor interoperability – so that both together aim at developing interoperability in the two major modalities of ordinary human experience)*. Moreover, the virtual is the fundamental ontological modality necessary for the natural description of phenomena or situations that can be explained in a more analytical, or more scientific, way in terms of interoperability – in the same way as virtual images are the phenomenological description of what could otherwise be described in terms of light rays and the laws of reflection and refraction. As a result, it is on a technical and psychological rather than philosophical basis that we can state the agent and its mentalistic attributes are virtual.

"Cognitive" interoperability (with quotation marks) between artificial agents is a purely technical notion; although it was not formulated in such terms, one can consider that "cognitive" interoperability was already the aim of the famous 1990 KSE (Knowledge Sharing Effort) project. To underline its roots in classical computer science, let us define it briefly, in a bottom up way, as the top of an ascending scale of abstraction levels:
- the lower levels of physical compatibility between computers and network equipments;
- the level(s) of network protocols and data exchange;
- the clasical level of data and programs interoperability: databases interoperability (with the two aspects of syntactic normalization – SQL – and "semantic integration" of databases schemas) and objects interoperability (through norms such as CORBA or Java RMI);

---

smarter and smarter". More than six years later, has anybody seen this invasion or this explosion? (This is not to deny the usefulness of CYC for practical purposes).

165 One might object that this is a very restrictive conception of AI. But we have not yet seen any AI product that could be considered as displaying "general intelligence" independent of any restriction on its operating context (even learning always occurs in predefined conditions). The notion of an operation framework seems to be essential to AI.

- the level of "cognitive" interoperability between artificial agents (through conformance to KSE or to the more recent FIPA specifications – which includes: normalized means for translating between knowledge representation languages, for referring to ontologies and for communicating by standardized messages).

The next step, semiotico-cognitive interoperability (without quotation marks) between Man and artificial agents, is a non obvious extension of the previous ones; it involves some form of communication between man and the machine, preferentially close to natural language; it may display different degrees. It means that, in its pre-defined semiotico-cognitive operation framework, the artificial agent appears to behave in the same way as a human agent would in the same situation and, in particular, that (to a predefined extent) some meanings *seem* to be shared between the user and the agent. Due to the restriction on the operation framework, this does not imply a positive answer to another sulfurous question: has the Turing test been passed? This means no more and no less than the agent has been designed properly, relative to its intended goals and operation framework as an agent. Semiotico-cognitive interoperability should therefore not be construed as a general *a priori* property of an AI system but as a *regulating goal* of AI systems development.

Interoperability is the technical notion that enables us to get rid of the identification paradigm (in which man and the machine are confused in some vicious circle) and invites us to think our relationship to the computer and to virtual agents rather than thinking ourselves as computers – as has long been the case in some AI or cognitive science circles (Berthier, 2005b).

Weaker than weak AI, stronger than Dennett's "intentional stance"

Differences with classical interpretations of AI can be stated: while strong AI is simply defused, our conception appears to be weaker than weak AI (or functionnalism) but stronger than Dennett's intentional stance.

According to our view of AI, strong AI is simply defused. Once a new ontological modality has been defined and it is shown appropriate to qualify the artificial agents of AI and their mentalistic attributes, asking if

these are "real" is essentially pointless: would anyone care to ask if a virtual image is real? That would be like trying to lit a fire with a virtual image. The scope of this remark should not be underestimated, considering that, disguised in new clothes, strong AI ideas are still lurking behind many undue interpretations or overstatements of some cognitive sciences results (for instance in relationship with the ideas of symbol grounding or embodiement).

But our conception is also weaker than weak AI or functionnalism[166]: instead of considering all observable behaviours of an artificial agent (observable by any objective external means), we are only interested in such behaviours as are observable by a human subject in a situation of interoperability with the agent, in its predefined operation framework. This includes two complementary restrictions: one on the artificial agent and its limited operation framework; the other on the observer, which must be human and in the appropriate situation. In the same way as you can perceive a reflection in a mirror only if you are at the proper place, you can perceive intentionality (or intelligence or any mentalistic attribute) in an artificial agent only if you interoperate with it in the proper operation framework. And you don't mind if a fly does not perceive it.

Morever, in both situations, it is a fact that you have no choice: even if you know that what you are seeing is just a reflection, you cannot avoid seeing it; even if you know that what you are interacting with is just a software component, you cannot avoid understanding its behaviour in terms of folk psychology: beliefs, goals, intentions and so on; this is in full accordance with the way we have defined the virtual. This distinguishes us from Dennett's intentional stance (Dennett 1987). Dennett's vocabulary (words in italics afterwards) may suggest that we have a choice: the intentional *stance* is the *strategy* that consists in *interpreting* the agent's behaviour *as if* it was a rational agent whose actions are determined by beliefs and desires. This is undoubtedly very relevant from the point of view of system developement (which is not the one we are talking about here), but it is much too weak to describe the unsophisticated user's natural attitude.

166 There are many conceptions of functionnalism (Pacherie,1993), but this is pointless here.

Naturalizing intentionality

Another consequence of our redefinition of the general modality of the virtual has been elicited in (Berthier, 2005c). It raises a new challenge for the cognitive sciences project of naturalizing intentionality[167]. Since we have shown that real and virtual intentionality must be distinguished, any "natural" explanation of intentionality should be able to make a distinction between these two modalities. But this means not only that it should explain intentionality on the basis of (biological, phylogenetical…) specificities of mankind, but also that it should do so in a way that cannot be applied to artifacts simulating these specificities. As a result, it is very difficult to imagine how general approaches, like Thom and Petitot's morphodynamics, could do the job. But it is no less difficult to imagine how approaches based on Varela's general idea of the embodiement of the mind could work by simply producing specialized levels of detail for this embodiement.

Information systems and business processes modelling

Practical consequences are not limited to AI. Consider the information systems (IS) domain. IS are a main factor of economic competition and they are tracking us in almost every part of our social lives. They have been rapidly changing in nature in the last decade: whereas they were traditionally considered as supports for the operational activities of an organization and they were essentially procedural, they now tend to be considered as supports for new kinds of activities (project management, decision making, innovation processes) that are highly interactive and very difficult to model in a procedural paradigm. Moreover, integrative IS (IS that can federate different organizations or different processes in one organization, through some kind of interoperability) and flexible or adaptive IS (IS that can adapt "easily" to the changing business processes in a rapidly evolving world) are becoming main themes for research.

---

167 Here, we can consider any of the variants of this project (for a review of these variants, see Pacherie, 1993).

Considering agents and associated concepts seems to be inescapable in the highly interactive and evolving IS of the future. At the technical level, multi agent systems have naturally been considered by many researchers as a possible answer to the above problems (Kishore & al., 2004; Wagner, 2003). But the same conclusion can be reached if we approach IS in terms that are meaningful for the organization it is designed to operate in and that allow one to make a link with its strategic orientations (Berthier & al., 2005).

In fact, numerous differences can appear between an IS such as it was initially specified and designed and its final software implementation; moreover, after its introduction in an organization, numerous factors may have changed: in the organization, in its human actors, in its environment, in the development or maintenance team or in the system itself. So that, finally, the rationality that had initially led to the design of the IS can have become globally opaque to anybody in the organization. As a consequence, there is a risk that the IS tends to: a) locally, constrain every human actor by the interoperability modes it impones on him; b) globally carry a kind of autonomized rationality, becoming a kind of agent whose implicit micro-decisions elude every possible control – even if one does not want to consider it as such. With some exaggeration, it appears as if there was an army of ghost agents that secretely maintained alive past forgotten decisions of the designers. Explicitly introducing agents in the design would make all this more explicit and easier to change.

Nevertheless, and this is an instance of the dilemma we started with, the business world displays some blocking. The notion of an agent inevitably evokes a variety of other highly mentalistic notions linked to the characteristics of agenthood. Most practionners of IS are not likely to unconditionally admit introducing concepts such as an artificial agent, its goals and intentions, nor are they ready to hear about "social conventions" that would tie together such agents and real human agents. Stating clearly from the beginning that all such concepts are meant as virtual could make things much easier. Of course, we do not mean that this would solve the IS problems we started with. There remains a lot of work to be done; for instance: what ethic rules and legal regulations should apply to agent based IS? But having clarified the ontological status of the artificial agents should help tackling such questions.

Conclusion

In order to solve the persistent dilemna of AI about the mentalistic attributes of artificial agents, this paper has introduced a new core meaning for the word "virtual", a meaning that distinguishes it clearly from "potential" – in opposition to all philosohical traditions but in accordance with many technical as yet isolated examples. As long as a software component appears to us as displaying some mentalistic attributes (intelligence, goal oriented behaviour, intentionality,…) in some operation framework, but can be denied these attributes in other contexts, then these attributes cannot be real; they have to be said virtual; and the agent itself must be said virtual. The new perspective on AI this definition allows has been related to its classical interpretations. Some practical consequences have been pointed out.

References

Bergson, H. (1896/1970). *Matière et mémoire*. trad. fr. in Œuvres, PUF, Paris (also available on the Web).

Berthier D. (2002). *Le savoir et l'ordinateur*. L'Harmattan, Paris.

Berthier D. (2004). *Méditations sur le réel et le virtuel*. L'Harmattan, coll. « Impacts des Nouvelles Technologies », Paris.

Berthier D. (2005a). Virtuel / virtual, *DITL, International Dictionary of Literary Terms*, J.M. Grassin ed., http://www.ditl.info/arttest/art6573.php

Berthier D. (2005b). Penser notre relation à la Machine, plutôt que nous penser comme des Machines, *Terminal*, n° 92, pp. 155-170.

Berthier D. (2005c). L'intentionnalité et le virtuel, *Intellectica*, n° 40, pp. 91-108.

Berthier D., Morley C. & Maurice-Demourioux M. (2005). Enrichissement de la modélisation des processus métiers par le paradigme des systèmes multi agents, *Systèmes d'information et management*, Vol. 3 n° 10, pp. 25-45.

Deleuze G. (1968). *Différence et répétition*. Presses Universitaires de France, Paris.

Dennett D. (1987). *The intentional stance*. MIT Press, A Bradford Book, Cambridge, MA.

FIPA: http://www.fipa.org

Granger G-G. (1995). *Le probable, le possible et le virtuel*. Odile Jacob, Paris.

Husserl E. (1931/1994). *Méditations cartésiennes*. Armand Colin, PUF, Paris.

Iglesias C., Garijo M., Centeno-Gonzalez J. & Velasco J., Analysis and Design of Multiagent Systems Using MAS-Common KADS, *Proceedings of the 4th International Workshop on Intelligent Agents IV, Agent Theories, Architectures and Languages*, Lecture Notes in Computer Science, Vol. 1365, Springer, 1997.

Jacob P. (2004). *L'intentionnalité*. Odile Jacob, Paris.

Kishore R., Zhang H. & Ramesh R. (2004). Enterprise integration using the agent paradigm : foundations of multi-agent-based integrative business information systems, *Decison Support Systems*.

Newell A. (1982). The Knowledge Level, *Artificial Intelligence*, Vol 59, pp 87-127.

Pacherie E. (1993). *Naturaliser l'intentionnalité*. PUF, Paris.

Schreiber G., Wielinga B. & Breuker J. (1993). *KADS: A Principled Approach to Knowledge-Based System Development*, Academic Press.

Wagner G. (2003). The Agent-Object-Relationship Metamodel : towards a unified view of state and behavior, *Information Systems*, n° 28, pp. 475-504.

Wooldridge M., Jennings N. & Kinny D., The Gaia Methodology for Agent-Oriented Analysis and Design, *Autonomous Agents and Multi-Agent Systems*, vol. 3, n° 3, p. 285-312, 2000.

COMPONENTIAL EXPLANATION IN PHILOSOPHY, COMPUTER SCIENCE, ORGANISATION THEORY AND AI

***Tibor Bosse*[1], *Catholijn M. Jonker*[2], *and Jan Treur*[1,3]**

[1]*Department of Artificial Intelligence, Vrije Universiteit Amsterdam*

[2]*Nijmegen Institute for Cognition and Information, Radboud Universiteit Nijmegen*

[3]*Department of Philosophy, Utrecht University*

tbosse@cs.vu.nl, C.Jonker@nici.ru.nl, treur@cs.vu.nl

Abstract. This paper shows how Componential Explanation as discussed within Philosophy relates to Compositional Verification in Computer Science and Artificial Intelligence. It is shown how a formal approach to Compositional Verification and some of the formal techniques developed for Computer Science and Artificial Intelligence can provide a formal basis and automated support for the notion of Componential Explanation as proposed in Philosophy. The role of interlevel relations is shown to be crucial in the formal analysis on which a componential explanation rests. Within application disciplines such as Biology, Cognitive Science and Organisation Theory, the importance of such interlevel relations is recognized as well. A case study has been undertaken to show the thoroughness of the approach and the level of detail needed to come up with a formal analysis that can serve as the basis of a componential explanation.

Introduction

The notion of componential explanation plays a role in different disciplines such as Philosophy, Biology, Cognitive Science, Organisation Theory, Computer Science and AI. Roughly spoken, *componential explanation* describes how properties of a system that is organised according to a number of components, can be explained from properties

of the components and their interactions. For componential explanation, Clark (1997) draws the analogy with modelling and analysis methods within AI, referring to, among others, Newell and Simon (1972) and Dennett (1978).[168] He also claims that componential explanation has a role to play in less classical AI areas such as connectionist approaches: in advanced connectionist work, complex tasks require highly structured multi-layer networks.[169] Clark (1997) gives suggestions, but does not address in more detail how to formalise componential explanation. This is the subject of the current paper. To this end methods developed originally in Computer Science are considered.

The area within Computer Science in which properties of component-based systems are analysed in terms of properties of their components is called *compositional verification*; e.g., Roever et al. (1998, 2001), Jonker and Treur (2002a). Formalisation and automation are important in the contributions to this area. The considered (software and hardware) systems are assumed to be hierarchically structured according to a number of aggregation levels. A central role is played by *interlevel relations* between properties at different levels of aggregation. For example, for a system S with property G that consists of two components A and B that have properties DP1 and DP2, respectively, the implication DP1 & DP2 & T $\Rightarrow$ G is an example of an interlevel relation expressing that S has property G in virtue of connectivity T and properties DP1 and DP2 of components A and B. Here the connectivity property T denotes a property that describes the connection or interaction between the components: transfers between the components. Compositional verification analyses properties of systems based on such interlevel relations.

In this paper it is explored how the notion of compositional verification developed within Computer Science relates to the notion of componential explanation as developed in an informal or semiformal sense within Philosophy, and the application disciplines Biology, Cognitive Science and Organisation Theory (cf. Cummins, 1975, 1983; Clark, 1997; Davies, 2001; Lomi and Larsen, 2001), and how it can be used to obtain a formalisation of componential explanation in a more technical sense, opening doors to existing or new software tools to support the explanation process. First the notion of componential explanation is briefly described (Section 2). In Section 3 it is discussed how the interplay of components at different aggregation levels and interlevel relations

168 'Modular programming methods in classical AI lent themselves quite nicely to a componential form of explanation. In attempting to understand the success of such a program, it is often fruitful to isolate the various subroutines, modules, etc. and to display their role in dividing the target problem into a manageable series of subproblems.' Clark, (1997, pp. 104-105)

169 'In such cases it is possible to advance our understanding of how the system succeeds by asking after the roles of these gross components (layers and subnets).' Clark, (1997, p. 105)

between them are considered important challenges for the application discipline Organisation Theory within the Social Sciences. In Section 4, compositional verification within Computer Science and AI is summarised. In Section 5, a case study is discussed in which the circulatory system is modelled from an organisational perspective. Based on this case study, Section 6 and 7 show how the different notions discussed (interlevel relations, componential explanation, and compositional verification) relate to each other. Section 8 concludes the paper with a discussion.

Componential Explanation in Philosophy

Hempel (1959) and Nagel (1961) focus on functional explanations why certain items I (such as the heart) are present within an organised system S (e.g., a human being). They base the explanation on an attempted form of deduction, concluding that the item I is necessary in the context of the overall system S (for a certain function F). In this line of reasoning the existence of functional equivalents is problematic: why would another item I' with the same functional contribution F not be possible instead? The dilemma is that:

- either functional equivalents exist, then the necessity of the existence of an item cannot be claimed deductively,
- or the necessity of the existence of an item can be claimed deductively, but functional equivalents are not allowed.

Hempel (1959) takes the first horn of this dilemma, Nagel (1961) the second one. Hempel's explanation does not provide a deductive argument. Nagel's is deductive, but requires a premise excluding the existence of functional equivalents, which is problematic (since there are no laws to derive it).

Cummins (1975) avoided this dilemma by a change of perspective. Instead of attempting to obtain a deduction concluding the existence of a certain item I, his deductive analysis A aims at concluding the systemic capacity C of the overall system S, on the basis of properties of the components of S. Within this analysis A the item I contributes function F. This function F is needed in A in the sense that, if it would be left out of A, capacity C cannot be deductively concluded anymore. Davies (2001, Chapter 2, pp. 25-27), discusses Cummins' account on componential explanation, also called *systemic functional analysis*; see also Clark (1997, Ch. 6). The idea is as follows. For a system S, one of its capacities C can be analysed: by virtue of what does S exercise C? For example, the capacity C of an animal to stay alive can be analysed in terms of different components within the animal and the jobs they perform: e.g., circulation, digestion, respiration.

According to Davies' analysis, first the subsystems performing such jobs are identified, and the relevant capacities specified. For example, within Biology the circulatory system contributes to C by a capacity C' to transport oxygen and nutrients to the places within the animal where they are used. A next level of functional analysis focusses on a capacity of one of these subsystems, for example the capacity C' of the circulatory system. Considering the next level, the analytical approach also needs to be performed for this subsystem, i.e., identification of the main components and the jobs they perform. Example capacities for this system are assimilation of oxygen and nutrients in the blood, propulsion of blood, and absorption of oxygen and nutrients. The heart is one of the contributing components for these capacities; in the context of capacity C' it can be attributed the (systemic) function F of pumping blood. After presenting a brief overview of Cummins' account, Davies (2001, Chapter 4) presents his own account on componential explanation. A main addition is that the phenomena analysed are *hierarchically organised*:

Let A denote the analysis of system S into its components, and C the systemic capacity analysed. The item I within S has systemic capacity function F if and only if:

(i*) I is capable of doing F

(ii*) A appropriately and adequately accounts for S's capacity to C in terms of the organised structural or interactive capacities of components at some lower level of organisation

(iii*) I is among the lower-level components cited in A that structurally or interactively contribute to the exercise of C

(iv*) A accounts for S's capacity to C, in part, by appealing to the capacity of I to F

(v*) A specifies the physical mechanisms in S that instantiate the systemic capacities itemised

Here (i*), (iv*), and (v*) are items of Cummins' account, and (ii*) and (iii*) are adding hierarchical organisation. Clark (1997)[170] considers componential explanation ('from parts to wholes', pp. 103-105) as a major explanatory strategy, to be used in conjunction with other types of explanation (for example, based on reciprocal input thought-action cycles, pp. 105-106), to explain interaction with the environment.

Componential Perspective on Organisation Modelling

The inherent complexity of the dynamics of multiple interacting processes within a society can be made manageable by organisation (Mintzberg, 1979, Kreitner and Kunicki, 2001). By using multi-agent organisation modelling techniques for analysis and simulation, this can be formalised;

---

170 '(1) An account of the gross behaviors of the well-functioning organism in the environment - an account that may invoke collective variables whose componential roots span brain, body, and world. (2) An account that identifies the various components whose collective properties are targeted by the explanations proper to (1). Two important subtasks here are to identify relevant neural components and to account for how these components interact. (3) An account of the varying information-processing roles played by the components (both internal and external) identified in (2) – an account that may well assign specific computational roles and representational capacities to distinct neural subsystems'. Clark (1997, p. 126)

e.g., (Lomi and Larsen, 2001; Ferber and Gutknecht, 1998; Ferber et al., 2001). In Nature, many phenomena have the same characteristic: they also involve complex dynamics of multiple distributed processes and their interaction. Therefore, a natural question is whether a multi-agent-organisation modelling perspective is promising for this domain of biological complexity.

Organisations can be viewed in two ways: (1) as adaptive complex information processing systems of (boundedly) rational agents, and (2) as tools for control. Central issues are (Lomi and Larsen, 2001):

- How to identify properties of the whole, given properties of parts; from the first view: 'given a set of assumptions about (different forms of) individual behaviour, how can the aggregate properties of a system be determined (or predicted) that are generated by the repeated interaction among those individual units?'
- How to identify properties of parts, given desired or required properties of the whole; from the second view: 'given observable regularities in the behaviour of a composite system, which rules and procedures - if adopted by the individual units - induce and sustain these regularities?'

Recently a number of formal and computational modelling techniques have been developed that can be used for simulation or for formal analysis of the dynamics within a multi-agent organisation. Examples of this formalisation trend can be found in books such as (Lomi and Larsen, 2001), and in a recently created journal: Computational and Mathematical Organisation Theory; e.g., (Moss et al., 1998). For an organisation, different levels of aggregation can be identified, from single agent behaviour to the dynamics of the overall organisation. Dynamics can be described in an abstract manner by focusing on one of these levels and specifying dynamic properties for this level. Moreover, interlevel relationships between dynamic properties at different levels can be identified.

One of the organisation modelling approaches that have been developed within the agent systems area is the Agent-Group-Role (AGR) approach, introduced in (Ferber and Gutknecht, 1998), and extended with a modelling approach for dynamic properties in (Ferber et al., 2001). According to this approach, the *organisational structure* is the specification of a specific multi-agent organisation based on a definition of groups, roles and their relationships within the organisation:

- An organisation as a whole is composed of a number of *groups*.
- A group structure identifies the *roles* and (*intragroup*) *interaction between roles*, and *transfers* between roles needed for such interactions.
- In addition, *intergroup* role relations between roles of different groups specify the connectivity of groups within an organisation.

For each of these elements both the structural aspect is specified and the dynamics/behaviour aspect (cf. Ferber et al., 2001). Thus three aggregation levels are considered: role, group, and organisation as a whole. Interlevel relations indicate how the specification at one level relates to this of an adjacent level; cf. (Ferber et al., 2001; Jonker and Treur, 2002b). In this way Lomi and Larsen (2001)'s challenges discussed above are modelled, from a perspective of componential explanation. The modelling approach is further explained and illustrated by the application to the circulatory system in mammals.

Compositional Verification

In this paper the formalization of Jonker and Treur (2002a) of compositional verification for Computer Science and Artificial Intelligence, summarized in this section, is used as starting point for the formalization of componential explanation. Within software engineering, the purpose of verification is to prove that, under a certain set of assumptions, a system will adhere to a certain set of properties, for example the design requirements. In this approach, verification is accomplished by a formal analysis of relations between properties and assumptions that respects the levels of aggregation already present in the compositional structure of the system.

A component-based system can be viewed at different levels of aggregation. Viewed from the top level, denoted by L_0, the complete system is one component S. At the next lower level of aggregation, level L_1, the system component S is a composition of components, and connections between these sub-components. Each component is again composed of its sub-components, and so on, until the lowest level of aggregation is reached, in which components are no longer composed of other components: primitive components.

The primitive components can be verified using dedicated verification methods, such as described in, e.g., (Leemans, Treur and Willems, 2002). Verification of a composed component is done using properties of the sub-components it embeds, and environmental properties of the component (i.e., assumptions on its embedding in the rest of the system). Given a set of environmental properties, the proof that a certain component adheres to a set of properties depends on the properties of its sub-components, and properties of the interactions between those sub-components. The compositional verification method can be formulated in more detail as follows:

A. Verifying one Aggregation Level Against the Other

- Determine which properties are of interest (for the higher level).
- Determine which assumptions (for the lower level) and which environment properties guarantee the higher-level properties.

- Prove the higher-level properties on the basis of these assumptions for the lower level and environmental properties.

B. The Overall Verification Process

- Determine the properties that are desired for the whole system.
- Apply procedure **A** iteratively until primitive components are reached.
- Verify the primitive components using techniques specialised for the type of component.

The results of verification are a hierarchy of properties at the different aggregation levels, and the logical relations between the properties of different aggregation levels, see Figure 1. In the picture, $P^j_{t,m}$ is the set of properties or assumptions of a component labeled j belonging to aggregation level L_t. This set is used in the proof for a component labeled m that is part of aggregation level L_{t-1}. Let $P_{t,m} = \cup_j P^j_{t,m}$. Then, the hierarchy is constructed such that $P_{t,m} \Rightarrow P^m_{t-1,r}$, for some r, being the label of the parent component of m on level L_{t-2}.

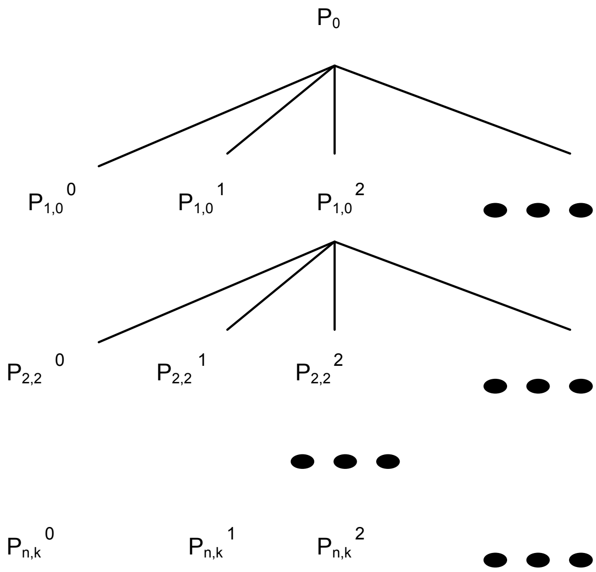


Figure 1. Hierarchy of properties for compositional verification.

Case Study: the Circulatory System

In this section, a case study in the domain of the circulatory system in mammals is used to illustrate how the philosophical idea of componential explanation can be worked out using the methods in compositional verification within Computer Science. This case study is often used as an example in philosophical literature. The analysis of the system's capacities in the case study is described in terms of *dynamic properties*: temporal statements that relate different states of a system (at different

time points) to each other. Such dynamic properties are identified at different aggregation levels. Next, *interlevel relations* are established, relating dynamic properties at different levels to each other. The properties have been formalised using the Temporal Trace Language TTL introduced in Jonker and Treur (2002a) (see also Bosse et al., 2006); for reasons of readability most of them are presented here in semiformal form. It is shown how this analysis can be used to obtain a componential explanation according to Cummins' and Davies' perspective.

The circulatory system takes care of a number of capacities, such as providing nutrients and oxygen to the body and taking waste (e.g., CO_2) out of the body; e.g., Noordergraaf (1978), Rideout (1991). The main property to focus on in this example is that the system provides oxygen for all parts of the body. The organisation of the circulatory system S is analysed as consisting of the following active components that (by showing their specific behaviours) all play their roles within the overall process: heart, capillaries in lungs and other organs, arteries (pulmonary artery channels, from the heart to the capillaries in the lungs; aorta channels, from heart to the capillaries in the body), veins (pulmonary veins, from the capillaries in the lungs to the heart; inferior and superior vena cava, from the capillaries in the body to the heart).

In Bosse et al. (2004), the circulatory system is modelled from an organizational perspective, following the AGR organisation modelling approach (Ferber and Gutknecht, 1998; Ferber et al., 2001). Following this approach, at the top level the system can be seen as one component. At lower levels, properties of sub-components (or *groups*) can be identified, as well as properties of transfers between these groups. The lowest level comprises properties of primitive components (or *roles*) and transfers between them. See Figure 2: at the top level, the circulatory system can be seen as one organization, which consists of two groups at a lower level, i.e., a *Pulmonary Cycle Group* and a *Systemic Cycle Group*. The main function of the Pulmonary Cycle Group is uptake of oxygen from the environment through the lungs, and the main the function of the Systemic Cycle Group is to supply this oxygen to the other organs. At the lowest level, each group consists of a number of roles with transfers between them. Note that both groups are organised according to a similar structure, consisting of the following five roles: *well*, *supply guidance*, *exchange*, *drain guidance*, *drain*.

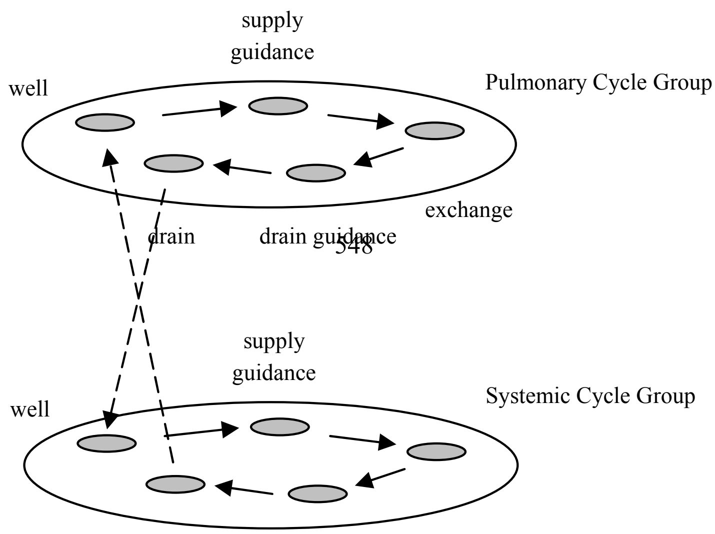


Figure 2. Roles, transfers within groups, groups and group interaction structures.

Moreover, to each role a certain active component (or agent) can be allocated. To be specific, for the Systemic Cycle Group, the allocation of agents to roles is as follows:

heart - systemic cycle well
aorta channels - systemic cycle supply guidance
organ capillaries - systemic cycle exchange
inferior and superior vena cava
- systemic cycle drain guidance
heart - systemic cycle drain

For the pulmonary cycle group instance the allocation of agents to roles is as follows:

heart - pulmonary cycle well
pulmonary channels - pulmonary cycle supply guidance
lung capillaries - pulmonary cycle exchange
pulmonary veins - pulmonary cycle drain guidance
heart - pulmonary cycle drain

Note that in both groups, the heart plays two roles, one of a well, initiating the flow, and one of a drain, where the flow disappears (and will re-appear in the other side). For more details about the model, see (Bosse et al., 2004).

In addition to this model, Bosse et al. (2004) present a number of dynamic properties relevant for the analysis of the system's capacities. In particular, the following properties are shown (all related to oxygen supply):

- environmental assumptions
- dynamic properties specifying component capacities
- dynamic properties for interaction between components (transfers)

These dynamics properties have been formalised using the predicate logic Temporal Trace Language TTL (cf. Jonker and Treur, 2002a; Bosse et al., 2006). Some examples of dynamic properties of the circulatory system (for reasons of presentation in a semiformal notation) are the following:

GP1(w) **Well successfulness (with maximal interval w)**

After an initiation time t0, for any point t there exists a time point t' with t < t' ≤ t + w such that at t' a fluid with ingredients I is generated by the well.

EA2(i) **Stimulus occurrence (with maximal interval i)**

For any point in time t there exists a time point with t < t' ≤ t + i such that at t' a stimulus occurs.

IrRI(c, r) **Drain– well intergroup role interaction**

At any point in time t0

if at some t ≤ t0 the drain within some group instance G_i received a fluid volume V with ingredients I

and between t and t0 no stimulus occurred

and at t0 a stimulus occurs

then there exists a time point t1 with t0 + c ≤ t1 ≤ t0 + r such that at t1 the well within the other group instance G_j generates a fluid volume V with ingredients I

GR(u, v, u', v') **Group successfulness**

At any point in time t,

if at t the well generates a fluid volume V with ingredients I

then there exist time points t' ≤ t'' with t + u ≤ t' ≤ t + v and t + u' ≤ t'' ≤ t + v' such that at t' ingredient A is added to the environment and ingredient B taken from the environment

and at t'' the drain receives a fluid volume V with ingredients I - A + B

RB1(e1, f1) **Supply guidance effectiveness**

At any point in time t

if the supply guidance receives a fluid volume V with ingredients I

then there exists a time point t' with t + e1 ≤ t' ≤ t + f1 such that at t' it generates a fluid volume V with ingredients I

Interlevel Relations for the Case Study

The idea of specifying dynamic properties at different aggregation levels is that the dynamics of the whole componential system can be (logically) related to the dynamics of lower levels. At the highest level, the following interlevel relation (between level 0 and level 1) holds:

Init & GR(s) & GR(p) & IrRI(s) & IrRI(p) & EA2 ⇒ GP1(s)

Thus, global property GP1(s) is implied by the lower level properties. Or, in other words, in all situations in which properties Init, GR(s), GR(p), IrRI(s), IrRI(p), and EA2 hold, property GP1(s) also holds. Here, the arguments (s) and (p) indicate that the property holds, respectively, for the Systemic Cycle Group or the Pulmonary Cycle Group. In a similar manner, the following interlevel relation can be established between properties at level 1 and 2:

IaRI1(s) & IaRI2(s) & IaRI3(s) & TR4(s) ⇒ GR(s)

An overview of all interlevel relations that are related to global property GP1(s) is depicted graphically in Figure 3 (comparable to Figure 1). These interlevel relations have been automatically checked using the model checker SMV (http://www.cs.cmu.edu/~modelcheck/smv.html; see also McMillan, 1993). This analysis also proved that none of the antecedents can be left out; in particular, if the heart's effectiveness fails, then GP1(s) cannot be concluded.

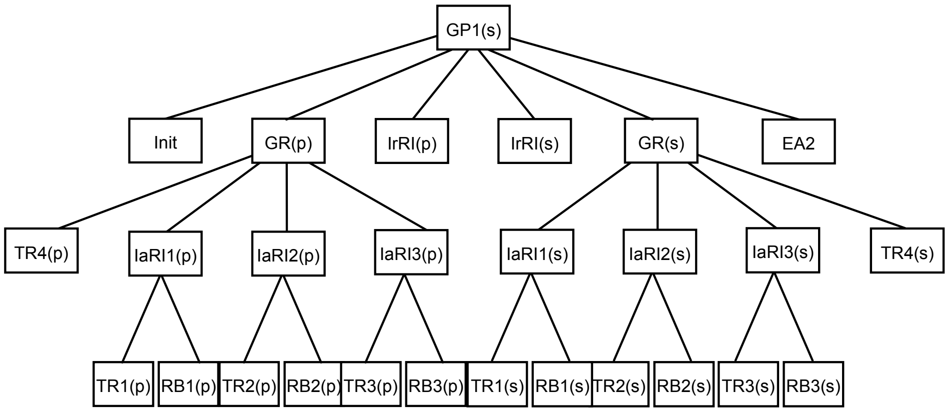


Figure 3. Interlevel Relations for Global Property GP1(s).

Componential Explanation for the Case Study

In the previous subsections a componential analysis A for the circulatory system S has been formalised by compositional verification methods from Computer Science. But to what extent does this indeed address componential explanation according to Cummins (1975, 1983) and Davies (2001)? As an example, consider the Aorta Channels as item I. The function F for this item is given by the property Supply Guidance Successfulness, RB1(s): if it receives a blood stream at one point, it will generate a comparable blood stream at another point. The system's capacity C is Well Successfulness GP1(s). Then the function Supply Guidance Successfulness within the system S is described by the following instantiated pattern according to Davies:

The item Aorta Channels within S has systemic capacity function Supply Guidance Successfulness if and only if:

(i*) The Aorta Channels satisfy Supply Guidance Successfulness

(ii*) The analysis appropriately and adequately accounts for S's capacity Well Successfulness in terms of the organised structural or interactive capacities of components at some lower level of organisation

(iii*) The Aorta Channels are among the lower-level components cited in the analysis that structurally or interactively contribute to the exercise of Well Successfulness

(iv*) The analysis accounts for S's capacity Well Successfulness, in part, by appealing to the capacity of the Aorta Channels to satisfy Supply Guidance Successfulness

(v*) The analysis specifies the physical mechanisms in S that instantiate the systemic capacities itemised

Indeed, (i*) to (iv*) are satisfied by the analysis above. However, to satisfy (v*), some specification of the physical mechanisms of the Aorta Channels has to be added, for example by referring to, e.g., Noordergraaf (1978).

Discussion

This article contributes to formalisation and automated support of componential explanation as developed in the area of Philosophy for application disciplines such as Biology, Cognitive Science and Organisation Theory. This is achieved by exploiting a formal framework of compositional verification as developed within Computer Science and AI. In particular, one of the formal approaches to compositional verification has been applied to a case study to provide a formal analysis, which can serve as the basis for a componential explanation that corresponds to the work of Davies (2001) and Cummins (1975, 1983). In addition, the article contributes to the area of Computer Science and Artificial Intelligence by making clear the conditions on componential explanation to bear on computer software, and provides an additional foundation for the ideas of Clark (1997), Dennett (1978), Newell and Simon (1972). The case study also shows the level of detail necessary to complete a formal analysis of only one aspect of the circulatory system that itself contributes to the capacity of an organism to live. The rigorousness of a formal approach to componential explanation therefore also begs for the development and use of dedicated software support. In the mean time, the formalization opens the doors to the use of existing tools that support verification in Computer Science, such as the model checker SMV.

The case study to analyse the circulatory system from an organisation modelling perspective has shown to be an appropriate example for the application of compositional verification. It may be expected that the approach is also applicable to other compositional systems (in disciplines such as Biology, Cognitive Science, and Organisation Theory).

References

Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A., and Treur, J. (2006). A Temporal Trace Language for the Formal Analysis of Dynamic Properties. Technical Report, Vrije Universiteit Amsterdam. http://www.few.vu.nl/~treur/TTL.pdf

Bosse, T., Jonker, C.M., and Treur, J. (2004). Organisation Modelling for the Dynamics of Complex Biological Processes. In: Lindemann, G., Moldt, D., and Paolucci, M. (eds.), Proc. of the International Workshop on Regulated Agent-Based Social Systems: Theories and Applications, RASTA'02. Lecture Notes in Artificial Intelligence, vol. 2934. Springer Verlag, pp. 92-112.

Clark, A. (1997). Being There: Putting Brain, Body and World Together Again. MIT Press, Cambridge, Mass.

Cummins, R. (1975). Functional Analysis. The Journal of Philosophy, vol. 72, pp. 741-760

Cummins, R. (1983). The Nature of Psychological Explanation, MIT Press, Cambridge, Mass.

Davies, P.S. (2001). Norms of Nature: Naturalism and the Nature of Functions. MIT Press, Cambridge, Mass.

Dennett, D. (1978). Brainstorms. MIT Press, Cambridge, Mass.

Ferber, J. and Gutknecht, O. (1998). A meta-model for the analysis and design of organisations in multi-agent systems. In: *Proceedings of the Third International Conference on Multi-Agent Systems* (ICMAS’98), IEEE Computer Society Press, pp. 128-135.

Ferber, J., Gutknecht, O., Jonker, C.M., Müller, J.P., and Treur, J., (2001). Organization Models and Behavioural Requirements Specification for Multi-Agent Systems. In: Y. Demazeau, F. Garijo (eds.), *Multi-Agent System Organisations. Proceedings of the 10th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'01*, 2001.

Fodor, J.A. (1983). *The Modularity of Mind,* Bradford Books, MIT Press: Cambridge, Massachusetts.

Hempel, C.G. (1959). The Logic of Functional Analysis. In: Gross, L. (ed.), Symposium on Sociological Theory; New York: Harper and Row, p. 271-287.

Jonker, C.M. and Treur, J. (2002a). Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness. In: (Roever et al., 1998), pp. 350-380. Extended version in: International Journal of Cooperative Information Systems, vol. 11, 2002, pp. 51-92.

Jonker, C.M., and Treur, J. (2002b) Relating Structure and Dynamics in an Organisation Model. In: J.S. Sichman, F. Bousquet, and P. Davidson (eds.), *Multi-Agent-Based Simulation II, Proc. of the Third International Workshop on Multi-Agent Based Simulation, MABS'02*, Lecture Notes in AI, vol. 2581, Springer Verlag, pp. 50-69.

Kreitner, R., and Kunicki, A. (2001). Organisational Behavior, McGraw – Hill

Leemans, N.E.M., Treur, J. and Willems, M. (2002). A Semantical Perspective on Verification of Knowledge. Data and Knowledge Engineering, vol. 40, pp. 33-70.

Lomi, A. and Larsen, E.R. (2001). Dynamics of Organizations: Computational Modeling and Organization Theories, AAAI Press, Menlo Park.

McMillan, K.L. (1993). Symbolic Model Checking: An Approach to the State Explosion Problem. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1992. Published by Kluwer Academic Publishers, 1993.

Mintzberg, H. (1979). The Structuring of Organisations, Prentice Hall, Englewood Cliffs, N.J.

Moss, S., Gaylard, H., Wallis, S., and Edmonds, B. (1998). SDML: A Multi-Agent Language for Organizational Modelling, *Comput. and Mathem. Org. Theory 4*, (1), 43-70.

Nagel, E. (1961). The Structure of Science. London: Routledge & Kegan Paul.

Newell, A. and Simon, H. (1972). Human Problem Solving. Prentice Hall.

Noordergraaf, A. (1978). Circulatory System Dynamics. Academic Press, New York.

Rideout, V.C. (1991). Mathematical and Computer Modelling of Physiological Systems. Prentice Hall, Englewood Cliffs.

Roever, W.-P. de, Langmaack, H., and Pnueli, A. (eds.) (1998). Proceedings of the International Workshop on Compositionality, COMPOS'97. Lecture Notes in Computer Science, vol. 1536, Springer Verlag.

Roever, W.-P. de, Boer, F. de, Hanneman, U., Hooman, J., Lakhnech, Y., Poel, M., and Zwiers, J. (2001). Concurrency verification: introduction to compositional and noncompositional methods. Cambridge University Press.

CAN MACHINES THINK? AN UPDATE AND REAPPRAISAL

Robert I Damper

School of Electronics and Computer Science

University of Southampton

+44 (0)23 80 594577

`rid@ecs.soton.ac.uk`

In 1950, motivated by his pioneering work in early digital computer technology, Alan Turing posed the question: “Can machines think?” and went on to offer the opinion that by the turn of the century, “one will be able to speak of machines thinking without expecting to be contradicted”. While artificial intelligence (AI) scientists and engineers have taken it as the ultimate challenge for their field to build a ‘thinking’ machine, philosophers have debated extensively the coherence of machine intelligence as a concept and the utility of the computational metaphor in understanding cognition. In no small measure, the history of attitudes to the question “Can machines think?” parallels the history of AI itself. The purpose of this paper then is to reappraise Turing’s question after more than 50 years of unprecedented technological advances in computing. Unfortunately, these technological advances have not generally been accompanied by increases in understanding of ‘intelligence’ and of the relations between minds and machines, so that Turing’s expectation of a positive answer to his question by the year 2000 has not been realized. An interesting issue is the extent to which thought is necessarily tied to consciousness. Hence, I finish with a brief appraisal of the current state of the scientific study of this phenomenon.

INTRODUCTION

Arguably, the seminal publication in artificial intelligence (AI) and cognitive science was Alan Turing's "Computing machinery and intelligence", which appeared in 1950. In his very first sentence, he writes: "I propose to consider the question, 'Can machines think'?" He went on to offer the opinion that by the turn of the century, "one will be able to speak of machines thinking without expecting to be contradicted" (p. 442). Few would disagree that this optimistic prediction remains unrealized. So what progress, if any, has been made towards reaching a definitive answer to Turing's question?

To many in the field, the search for thinking machine was and remains something of a holy grail for AI. For instance, Newell (1973, p. 25) offers the opinion that AI could as well be called *theoretical psychology*, although elsewhere (Newell, 1990) he writes "AI is a branch of computer science" raising the interesting issue of what he thinks is the relation between the two characterizations. Yet, as is well-known, in his 1950 paper, Turing very quickly abandons the "Can machines think?" form of the question as "… too meaningless to deserve discussion" because, he says, of the difficulties of definition of the words 'machine' and 'think'. Instead, he attempts "to replace the question by another" and moves to the description of an 'imitation game', which—in one of its forms—shortly came to be called the Turing test. However, perhaps not surprisingly as it involves replacement of one question by another (actually several, see p. 442), debate surrounding the Turing test is no less than that surrounding the original question.

The view expounded here is that the history of attitudes to Turing's famous question is virtually a proxy for the history of AI itself. Hence, after some brief scene-setting in which I seek to assess Turing's own attitude to his famous question, I look at how attitudes to it have varied during three main periods. The first (Section 3), covering the 1950's and early 60's, I loosely characterize as the period of British cybernetics. In this time frame, automatic computing was a nascent and esoteric activity, the capabilities of electronic computers were strictly limited, and the philosophical issues surrounding the potential and ultimate limits of computation were still being identified. There followed a spell of rapid technological advance (dealt with in Section 4), characterized as the heyday of GOFAI, in which optimism about the prospects for symbolic AI flourished. Thereafter, moving in to the 1980's and beyond (Section 5), this optimism waned largely as a result of failure to scale up some of GOFAI's successes with toy problems to sensibly large real problems. Symbolic AI underwent something of a decline, to be replaced by connectionism and so-called embodied AI. In many respects, we are no further forward in answering the question now than in 1950. In more recent years, however, a number of influential commentators and thinkers

have argued that a key issue was being forgotten in previous debate about the nature of thought and the prospects for replicating thought processes in artifacts—namely consciousness. Hence, I make some necessarily brief comments on this exceedingly complex and controversial issue before concluding.

TOO MEANINGLESS TO DESERVE DISCUSSION?

Did Turing really believe his question too meaningless to deserve discussion? Perhaps not, given that his clear motive must have been to introduce the Turing test as a way of sidestepping some of the extreme difficulty inherent in answering the original question. But, as Moor (1976) writes some years later:

> "... it is difficult to understand ... the propriety and adequacy of the replacement [i.e., the Turing test] if the question being replaced is too meaningless to deserve discussion."

During his 1951 BBC Radio broadcast, Turing said:

> "If now some particular machine can be described as a brain we have only *[!]* to programme our digital computer to imitate it and it will also be a brain." (cited by Copeland 2004, p. 478)

But given that a brain is patently an organ for thinking, this appears to stand in stark contrast to his earlier negative description of the main question as "… too meaningless to deserve discussion". Of course, Turing may just have been speaking counterfactually (*as if* a machine could be described as a brain!), but the fact that he was addressing a lay public, and the tone of the rest of the broadcast, encourages us to take him at face value. This much more optimistic view of the prospects for building thinking machines seems to be the one attributed to him by those who, from early contributions such as Wilkes (1953) and Mays (1953) through to the present day, have attacked the computational 'brain–machine' analogy and/or the Turing test as a useful indicator of machine intelligence.

Turing's reluctance to address the question directly apparently stems from the difficulty that he saw in defining the terms 'thinking' and 'machine'. So is the question merely semantic? As regards *thinking*, we might well ask if there are different forms: human and machine, for instance. In 1950, the idea of a 'thinking machine' was undoubtedly fairly radical, except perhaps to a very few initiates. Whereas debate had raged for some time concerning human versus animal 'thought', the possibility of machine thought was only entertained in restricted circles. Generally, machines were seen as the product of an explicit human design process (cf. the so-called Lovelace objection to machine intelligence), and since no one had a very sound idea what exactly

thought entailed, how could it be designed in to an artifact? Yet to Turing as a logician, as to Boole and Babbage before him, there was a strong relation between, at least, logical mathematical reasoning and the sort of "effective procedures" which could be mechanized. This led to the notion that the human brain might usefully be viewed in mechanical terms, as a 'machine', so opening up the way for the computational metaphor to dominate cognitive science for the next 50 years (and probably more).

THE 1950'S AND EARLY 1960'S

Turing's question was quickly taken up for debate by contemporary computer pioneers and philosophers. Thus, Maurice Wilkes addressed it in a 1951 BBC broadcast (in the same series as Turing's contribution) and also in a subsequent publication (Wilkes, 1953). Other notable publications of this period were those of Wolfe Mays (1953) and Mario Bunge (1956). In general, and in contrast to Turing himself, who was enthusiastic about the prospects for machine intelligence, commentators of this period were generally antagonistic to the notion that a machine could think 'for itself'.

Mays (1951, 1953) asserts that the question is not merely semantic but a matter of fact. For him, the answer is "yes" only if machine and humans reach their conclusions by similar processes. (But how would we know this?) He coins the term 'thinking by proxy'; a form of *as if* thinking. Noting that computing machines perform a kind of calculus, by execution of an algorithm, Mays writes "... a calculus is the very antithesis of thinking". This view is in sharp contrast to that of Turing himself, and to the later AI symbolists (e.g., Newell, 1980), who saw the very essence of intelligence as lying in performing myriad tiny steps, each of them mind-numbingly trivial by themselves, but adding up to something more than the sum of the parts. As an interesting aside, a part of Mays' argument is based on what we now recognize as the symbol grounding problem (Harnad, 1990). Mays writes: "if we grant that these machines *[i.e., digital computers]* are complex pieces of symbolism, … it is clear that in order to acquire a significance the symbols need to be linked with a set of referents" (p. 249).

By contrast to Mays, Bunge does seem to think Turing's question is merely semantic. To understand better his objections, let us sharpen our ideas about the semantics of questions like "Can machines *X*?" (where *X* is some activity in the natural world) by posing the apparently simplistic and innocuous question "Can airplanes fly?" To most readers, the answer is self-evidently yes. Thousands of planes fly routinely every day. Yet before the invention of the aircraft, the only extant example of heavier-than-air powered flight was offered by birds. Although there is some sense in which an airplane flies just as a bird does, the two are also

rather different in certain very obvious respects. So is there a proper analogy between a bird and an airplane 'flying' and between a human and a machine 'thinking'? While we understand the mechanics of flight reasonably well—well enough to see where the main differences between animal and machine flight might lie—there is not the same understanding of thinking to allow us to settle the matter. Now consider the question: "Can cars walk?" This one is obviously more problematic! But cars do get the occupants from *A* to *B*, *as if* they had walked. However, as Bunge (1956) writes:

> "… to assert that they *[machines]* think is as erroneous as saying that cars walk … This fallacy of inferring that something acting for us must … participate in human nature, is typical of primitive and archaic logic."

So is (artificial) 'thinking' semantically more like 'flying' or like 'walking'? How apposite is the metaphor? In the early days of computer science and technology, it was admittedly stretched. But as technology advances, does not the description fit better, much as these days we barely give a second thought to talk of "walking robots"?

In general then, the consensus during the 1950's and early 1960's (at least among fellow British scientists and philosophers) was to answer Turing's question in the negative. If there was such a thing as machine 'thought', it was so impoverished and different to human thought as to be barely worthy of the name. This consensus was strongly predicated on the view of computers as a tool for extending human computation ... not different in kind from pencil and paper, and requiring human supervision and interpretation (cf. Bunge's "something acting for us"). But, although Turing was arguably the first to coin the term 'machine intelligence', across the Atlantic AI was developing a more assertive character, in which the notion of the thinking computer was attracting widespread acceptance (cf. Raphael, 1976).

THE LATE 1960'S AND 1970'S

Over this time span, a change of opinion regarding Turing's question is discernible, as the capabilities of electronic computers start to grow and as the early influence of the British cyberneticists wanes to be replaced by ideas of machine functionalism (Putnam, 1967)—first in philosophy of mind and then in symbolic AI. Obviously, a complete history of AI over this burgeoning period is out of place here. Let us just remark that some landmarks in the development of AI at this time were Newell and Simon's GPS (see Newell, 1963), DENDRAL (Buchanan, Sutherland and Geirgenbaum, 1969), SHURDLU (Winograd, 1971), AM (Lenat, 1976), MYCIN (Shortliffe, 1976), and so on. Early successes, especially in the

realm of expert systems, seemed to confirm the potential of the symbolic approach (relative to connectionism as it had then developed) and led to great optimism for the future of artificial intelligence. This optimism culminated in the physical symbol system (PSS) hypothesis of Newell (1980), in which ideas of functionalism and implementation-independence were brought to their ultimate expression in the idea that formal symbol manipulation, as in a digital computer, was literally constitutive of intelligence.

This period was the heyday of what John Haugeland shortly after famously called Good Old Fashioned AI, or GOFAI (Haugeland, 1985). The view starts to emerge of machine 'thinking' as worthy of the description. To some, such as Newell and other adherents to the PSS hypothesis, there was no difference of kind between human and machine thinking, both being the product of formal symbol manipulation. To others (e.g., Moor, 1976), machine intelligence posited a valid but possibly different kind of 'thinking'.

THE 1980'S AND BEYOND

From the perspective of 2006, it is perhaps hard to understand that anyone ever took the PSS hypothesis seriously. I certainly find it hard to do so. It seems to miss the mark almost entirely in capturing the essence of intelligence. A straw poll of the students in my Artificial Intelligence lecture class, taken annually, reveals a declining number willing to say that they accept it as a reasonable characterization of intelligence. Few would disagree that there has been something of a demise of GOFAI. What happened to bring about this state of affairs? Arguably, there were three main factors: first, the spectacular resurgence of the connectionist brand of AI (witness the enormous impact of the back-propagation algorithm of Rummelhart, Hinton and Williams, 1986); second, a philosophical shift brought about by certain influential articles seeking to make explicit the flaws in symbolic AI; and, third, an increasing belief that the key to understanding thought must somehow be tied up with understanding consciousness, a topic largely ignored in earlier eras. It would not be appropriate to consider the large and diverse field of connectionism further here, and we will postpone our necessarily all-too-brief brief treatment of consciousness to a later section. But some remarks on the remaining one of these three issues are called for. Possibly foremost among the influential articles alluded to above are those of Rodney Brooks (1990, 1991), in which he laid the foundations of a new 'embodied' AI that minimized the role of internal (symbolic) representation and argued that interaction of an embodied agent with its external world is the cornerstone of intelligence.

Even prior to Brook's work, philosophical discussion of the matter was massively and notoriously influenced by John Searle's celebrated Chinese room argument (CRA), which appeared in 1980 and purported to show the futility of the search for 'strong AI'. The latter is, loosely speaking, Searle's term for the brand of GOFAI holding that "the appropriately programmed computer really is a mind". There is, of course, a vast literature—too large to review here—surrounding the CRA. Indeed, 15 years ago Gomilla (1991) described this literature as "nearly infinite" and it has continued to grow since! Almost all of it is aimed at denying Searle's negative conclusion in some way or other. (For a notable exception, see Maloney, 1987.) Although the majority opinion accepts the so-called 'systems reply' as a conclusive counter to the CRA, there is a bewildering variety of other rejoinders, and remarkably little consensus overall on exactly how Searle's argument is flawed (Damper, 2004). I am personally disposed to attacks on the logical form of the argument and/or the way this is derived from the informal (natural) language in which it was originally couched (Häggqvist, 1996; Damper, in print). But even given the wide-spread reluctance of commentators to accept Searle's point of view, still it seems he has had some influence in encouraging a more realistic view of AI's weaknesses to propagate. (At least, he himself claims this to be the case.)

In spite of Searle's opposition to the Turing conception of machine intelligence, he is forthright in believing that a machine *can* indeed think. To him, this is trivially true—on the (extraordinarily shaky) grounds that the brain actually *is* a machine. I do not see that this can be easily accepted as an obvious fact, but putting that aside, Searle's point seems to be that the brain is a 'machine' for generating consciousness by virtue of its having "the right causal powers"—whatever those might be. And lest one thinks that Searle's acceptance of the brain as a 'machine' somehow diminishes his own argument against machine intelligence, we should perhaps remember his oft-repeated (but far from universally accepted) warnings against confusing mere simulations with 'the real thing', and reflect that Turing was undoubtedly referring to simulation in the quote above, taken from his 1951 radio broadcast.

Since Searle's Chinese room argument turns on acceptance of the Turing test conception of intelligence, it is no surprise to see that there is also an enormous literature debating the pros and cons of this test. Some key works are those of Moor (1976), French (1990), Copeland (2000), Saygin, Cicekli, and Akman (2000) and Halpern (2006). Opinions differ on whether or not the Turing test was intended to be an operational definition of machine intelligence, or 'thinking'. In particular, and contrary to the assumptions of most commentators, Copeland (2000) points out that no such definition is to be found in Turing's 1950 paper and quotes Moor (1976) to the effect that the "… value of the imitation game lies not in treating it as … an operational definition but … as a potential source for

good inductive evidence for the hypothesis that machines think" (p. 249). But this seems to take us full circle back to the original question.

Thus far, things do not appear to have moved on greatly from 1950, and the very dawn of the computer age, in respect of answering the philosophical questions surrounding machine intelligence, cognitive science, the computational metaphor for mind, etc. Indeed, a quote from Tyler Burge is apposite:

> "In my view, the term [Cognitive Science] is best taken to be like "Christian Science" not only in denoting a doctrine rather than a discipline, but also in being a proper name rather than a description." (Burge 1982, p. 286)

THE MYSTERY OF CONSCIOUSNESS

In recent years, many prominent thinkers have pointed out the regrettable omission of the key issue of consciousness from discussions of AI (as 'synthetic psychology') just as from mainstream psychology. These contributions include Penrose (1989), Dennett (1991), Edelman (1992), Chalmers (1996) and others. The basic thesis is that thought by itself makes little sense divorced from consciousness. So is there any enlightenment to be gained by introducing this admittedly thorny issue into the debate?

In *The Mystery of Consciousness* (1997), Searle reviews and debates the contemporary work referred to above. My reading of the book, which I take to be uncontroversial, is that none of Penrose, Dennett, Edelman, etc. has the answer to the mystery. Searle's basic message is a call for reappraising the notion of reductionism: Consciousness is irreducible but this does not imply dualism. But in the absence of a concrete reappraisal, how far does this get us?

So is there any valid scientific basis to the study of putative machine intelligence? Or is it merely a label attached to a branch of engineering (much as the description is used in the title of the IEEE's *Transactions on Pattern Analysis and Machine Intelligence*) but devoid of meaningful relation to what goes on in the brains of sentient, conscious beings? Certainly, Brooks points to a lack of scientific basis when he writes:

> In my opinion we are completely prescientific … about what consciousness is. We do not know exactly what it would be about a robot that would convince us that it had consciousness, even simulated consciousness. (Brooks 2002, p. 194)

If we cannot yet expound a scientific foundation for the study of consciousness, we can at least indicate where the intellectual action currently lies. I will turn therefore to what appears to me to be the principal controversy in contemporary views of consciousness, namely

the importance (or otherwise) of *qualia*—the subjective phenomenon, or *feeling*, of experience. This takes us right back to Mays (1953) and his objection to Turing that a machine designed to perform precise mathematical and logical operations is a poor candidate for exhibiting "intuitive, often vague and imprecise, thought processes", since intuition appears (intuitively!) to be strongly tied to experience. While some philosophers, pre-eminently Chalmers (1996), take the explanation of qualia to be the "hard problem" of consciousness, AI scientists and engineers tend to deny or sidestep the issue. For instance, Baum (2004) argues that we have subjective experience (e.g., pain) just because it has evolutionary survival value, and writes: "You have a sensation of pain because these neurons here fire" (p. 68). To others (myself included), such crude appeals to identity theory (cf. Crick 1994) explain nothing; they are a sleight of hand—see Searle (1997, p. 30). Whether or not qualia turn out to be a key issue in understanding consciousness—natural, machine or both—we cannot yet say; the jury is still out and likely to be so for some long time to come.

CONCLUSIONS

Answering Turing's question "Can a machine think?" is virtually synonymous with the AI project. Arguably, cognitive science—by its acceptance of the 'brain as computer' metaphor—already assumes a positive answer and moves to consider the question: exactly how and what does the brain compute? Turing's optimism back in 1950 regarding at least pragmatic acceptance and usage of the term 'machine thought' in everyday conversation by the year 2000 has not come to fruition. In fact, one might well argue it was misplaced.

Looking at the 50-60 year history of AI, worrying cycles of optimism and pessimism in answering the question can be discerned. One would be hard pressed to argue that any great progress has been made. Attempts by philosophers and pioneer symbolists (like Putnam and Newell) in the 1960's and 1970's to gain acceptance for the idea that computers already could 'think' just as we do, that mental states are computational states and vice versa, made initial headway but then were increasingly seen as simplistic and devoid of supporting argumentation and evidence, almost unscientific in fact.

To many, it seemed that a key ingredient was self-evidently missing from these musings. Human thought appears inextricably tied to human consciousness, so can there be meaningful thought without consciousness? If not, the search for machine intelligence has to encompass a search for machine consciousness; and scientific progress in understanding consciousness is essential to this enterprise. Yet any such understanding is almost entirely absent ('prescientific' as Brooks

says) at present. Turing's question is an empirical and open one. At the moment we have no answer. Whether or not we can ever achieve machine intelligence, design thinking machines, we have no idea. And we have no real idea either how to make progress in providing an answer. We simply don't know enough.

REFERENCES

E. Baum (2004). *What is Thought?* MIT Press/Bradford Book, Cambridge, MA.

R. A. Brooks (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1): 3-15.

R. A. Brooks (1991). Intelligence without representation. *Artificial Intelligence*, 47(1-3): 139-159.

R. A. Brooks (2002). *Robot: The Future of Flesh and Machines*. Penguin, London, UK.

R. G. Buchanan, G. Sutherland and E. Geirgenbaum (1969). Heuristic DENDRAL: A program for generating explanatory hypotheses in analytical chemistry In B. Meltzer and D. Michie, *Machine Intelligence 5*, Edinburgh University Press, Edinburgh, UK.

M. Bunge (1956). Do computers think? (II). *British Journal for the Philosophy of Science*, 7(27): 212-219.

T. Burge (1982). Two thought experiments reviewed. *Notre Dame Journal of Formal Logic*, 23(3): 284-93.

D. Chalmers (1996). *The Conscious Mind*. Oxford University Press, New York, NY.

B. J. Copeland (2000). The Turing test. *Minds and Machines*, 10(4): 519-39.

B. J. Copeland, editor (2004). *The Essential Turing*. Oxford University Press, Oxford, UK.

F. Crick (1994). *The Astonishing Hypothesis*. Simon and Schuster, London, UK.

R. I. Damper (2004). The Chinese room argument: Dead but not yet buried. *Journal of Consciousness Studies*, 11(5-6): 159-169.

R. I. Damper. The logic of Searle's Chinese room argument. *Minds and Machines*, in print.

D. Dennett (1991). *Consciousness Explained*, Penguin, London, UK.

G. M. Edelman (1992). *Bright Air, Brilliant Fire: On the Matter of the Mind*, Basic Books, New York, NY.

R. M. French (1990). Subcognition and the limits of the Turing test. *Mind*, 99(393): 53-65.

A. Gomila (1991). What is a thought experiment? *Metaphilosophy*, 22(1-2): 84-92.

S. Häggqvist (1996). *Thought Experiments in Philosophy*. Almqvist & Wiksell, Stockholm, Sweden.

M. Halpern (2006). The trouble with the Turing test. *The New Atlantis Review*, Number 11, Winter 2006, 42-63. (Available at `http://www.thenewatlantis.com/archive/11/halpern.htm`.)

J. Haugeland (1985). *Artificial Intelligence: The Very Idea*. Bradford Books/MIT Press, Cambridge, MA.

S. Harnad (1990). The symbol grounding problem. *Physica D*, 42: 335-346.

D. B. Lenat. *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search*. PhD thesis, Stanford University, 1976.

J. C. Maloney (1987). The right stuff. *Synthese*, 70(3): 349-372.

W. Mays (1951). The hypothesis of cybernetics. *British Journal for the Philosophy of Science*, 2(7): 249-250.

W. Mays (1953). Can machines think? *Philosophy*, 27(2): 148-162.

J. H. Moor (1976). An analysis of the Turing test. *Philosophical Studies*, 30(4): 249-257.

A. Newell (1963). A guide to the General Problem-Solver program GPS-2-2. RAND Corporation, Santa Monica, CA. Technical Report No. RM-3337-PR.

A. Newell (1973). Artificial intelligence and the concept of mind. In R. C. Shank and K. M. Colby (Eds.), *Computer Models of Thought and Language*, Freeman, San Francisco, CA, 1-60.

A. Newell (1980). Physical symbol systems. *Cognitive Science*, 4(2): 135-183.

A. Newell (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

R. Penrose (1989). *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*. Oxford University Press, New York, NY.

H. Putnam (1967). *Psychological predicates*. In W. H. Capitan and D. D. Merrill (Eds.), *Art, Mind, and Religion*, University of Pittsburgh Press, Pittsburgh, PA, 37-48.
B. Raphael (1976). *The Thinking Computer*. W. H. Freeman, San Francisco, CA.
D. E. Rumelhart, G. E. Hinton and R. Williams (1986). Learning representations by back-propagating errors. *Nature*, 323(9): 533-536.
A. P. Saygin, I. Cicekli and A. Akman (2000). Turing test: 50 years later. *Minds and Machines*, 10(4): 463-518.
J. R. Searle (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3): 417-457 (including peer commentary).
J. R. Searle (1997). *The Mystery of Consciousness*. Granta, London, UK.
E. H. Shortliffe (1976), *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York, NY.
A. M. Turing (1950). Computing machinery and intelligence. *Mind*, 59(236): 433-460.
M. Wilkes (1953). Can machines think? *Proceedings of the Institution of Electronic and Radio Engineers*, 41: 1230.
T. Winograd (1972). *Understanding Natural Language*. Academic Press, London, UK.

Section 3

enquiry logic & thought

ON THE POSSIBILITY OF HYPERCOMPUTING SUPERTASKS

VINCENT C. MÜLLER
Dept. of Philosophy & Social Sciences

American College of Thessaloniki

+30-2310-398 211

vmueller@act.edu

This paper investigates whether hypercomputing forces a revision of the standard reading of the Church-Turing thesis. It discusses proposals for computing machines that are said to compute an infinite number of computing steps in finite time. It argues that these proposals fall into a dilemma: either they cannot be specified such that they have output states, or they compute with contradictory output states. Therefore, infinite hypercomputing is no reason to reject the standard reading of the Church-Turing thesis. The investigation of computing with supertasks also indicates that certain supertasks are impossible, including Zeno-type movements.

Introduction: Church-Turing and Hypercomputing

Copeland and the Church-Turing Thesis

The theoretical literature on the notion of computing, whether it is in the context of computationalism in the philosophy of mind, the possibility of artificial intelligence or mathematical computation, has traditionally assumed as background consensus that what a computer can do in principle is identical to what is "effectively computable", i.e. what can be computed by the mechanical application of a definite rule of finitely many instructions – of an algorithm.[171] The notion of computability was accordingly defined by Church, Turing and others in what is now known as the "Church-Turing thesis", one formulation of which is: *all and only the effectively computable functions can be computed by a Turing machine*. Under the traditional, strong, interpretation, this is thought to cover effective computation by both humans and machines. Strictly speaking, Church's thesis is that all effectively computable functions are recursive, and Turing's thesis is that all effectively computable functions are computable by the Turing-machine. Since the inversions to both theses are known to be true, to call a procedure "effective", "algorithmic", "recursive" or "Turing machine computable" all comes down to the same.[172]

In a series of papers, Jack Copeland and others have argued that the Church-Turing thesis has been widely misunderstood and actually expresses a weaker thesis, concerning what a *human* can effectively compute, thus concluding that the Church-Turing thesis says nothing about what is computable *by machines*, or computable *in principle*. All sorts of errors in the philosophy of computing and mind are blamed on this alleged misunderstanding.[173] If this were correct, one would have to

---

171 Indicative for the philosophy of mind: Fodor 2000, Scheutz 2002, Piccinini 2004, Churchland 2005, Fodor 2005, Pinker 2005; for artificial intelligence: Copeland 1993; for mathematical logic: Boolos, Burgess and Jeffrey 2003, ch. 3ff.

172 Church 1936, Turing 1936/1937; cf. Boolos, Burgess and Jeffrey 2003, ch. 3ff; Harel 2000.

The notion of "Turing machine" is well explained in many places, see particularly Penrose 1989, ch. 2; Floridi 1999, 26ff; Davis 2000, ch. 7; Copeland 2003, 4ff.

173 Copeland 1997, 1998, 2000, 2002a, 2002b, 2004; Copeland and Proudfoot 1999, 2000; cf. Shagrir and Pitowsky 2003.

Concerning the historical question what Church, Turing and other contemporaries had in mind, Copeland mentions that in the 1930ies and 40ies, the word "computer" meant a person doing computation. While it is true that universal computing machines did not exist before 1941 (the "Z3"), there had been non-electronic calculating machines for centuries and Turing was hardly unaware of the possibility of programmable (universal) computing machines. There are strong indications that Church (in his 1936 paper) and Turing thought the thesis to apply to machines, too. Turing states in the opening

distinguish one notion for both humans and machines ("effective", "Turing machine computable") and a wider one for machines only ("algorithmic", "recursive"). Copeland focuses only on the notion for machines and calls this part of the traditional strong interpretation of the Church-Turing thesis the "maximality thesis", stating it as follows: "all functions that can be generated by machines (working on finite input in accordance with a finite program of instructions) are Turing machine computable" (2000, 15). He says that while the Church-Turing thesis is true of humans, the maximality thesis is "known to be false" if we take the machines to be "machines in a possible world" (Copeland 2000, 15; cf. 31). "It is straightforward to describe abstract machines that generate functions that cannot be generated by the UTM" (2004, 12). What remains contentious on his view is merely whether the maximality thesis is true in the actual world.

To be sure, the set of all functions (even of all functions over the positive integers) is larger than the set of Turing-computable functions, since the latter is denumerable, while the former is not. But I will show that it is far from straightforward to "describe abstract machines" that *compute* such functions, while avoiding contradiction - which will shift the burden of proof to those who want to reject the traditional strong reading of the Church-Turing thesis.

Before we go into the details, let it be clear that the Church-Turing thesis concerns only *digital* or "discrete state" computing. This follows directly from the restriction to effective algorithmic procedures, which proceed step by step, where steps are distinguished by a discrete state. Whether non-digital, i.e. "analogue", or "continuous" computing deserves the name of "computing" and whether analogue mechanisms could compute functions that are not Turing-computable are matters not relevant to our point here. As Siegelman (Siegelman 1995, Siegelman and Sontag 1998) and many others (e.g. Bringsjord und Zenzen 2003) have shown, there is reason to believe that analogue mechanisms are possible which can compute functions that are not Turing-computable.

---

paragraph of his 1936 paper: "The 'computable' numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. … According to my definition, a number is computable if its decimal can be written down by a machine." About the concept from his 1936 paper that we now call a "Turing machine", he said in a 1947 address to the London Mathematical Society: "I considered a type of machine which had a central mechanism, and an infinite memory which was contained on an infinite tape…. One of my conclusions was that the idea of a 'rule of thumb' process and a 'machine process' were synonymous." (Turing 1992, 106). For this process, he rejects infinitely many digital states: (1936) §9 "If we admitted an infinity of states of mind, some of them will be 'arbitrarily close' and will be confused."

Finally, it would appear to be the point of Turing's 1936 paper to show that all effectively computable functions are computable by his machine, and thus that the halting problem of his machine is the Entscheidungsproblem. So it would be odd to have this problem for humans, but not for machines. (Odd and a dramatic inversion of the Penrose/Lucas argument!)

However, this would not refute the strong interpretation of the Church-Turing thesis. The situation would still be described very aptly by Floridi, when he says: “From Turing power up, computations are no longer describable by algorithms” (1999, 36). Accordingly, the strong Church-Turing thesis defended here does not imply Copeland’s “maximality thesis”, since the latter makes no mention of algorithms - an absence that is used by Copeland to attack it with the possibility of analogue computers (that the strong thesis is not *identical* to the maximality thesis is evident from the latter’s restriction to machines).

Hypercomputing

The rejection of the Church-Turing thesis under its strong interpretation is motivated by the idea that there could be machines that could compute what no human and no Turing machine could compute.[174] This computing of what is not Turing-machine computable is now called “hypercomputing”. Proposed designs include Turing’s “O-machines” (“oracle machines” with a black box that answers non-computable queries non-mechanically[175]), “Zeno machines” (that can compute infinitely many steps, see below), analogue computers (but see above), quantum computers, Putnam-Gold machines (computers that may “change their mind”), probabilistic machines, machines in Malament-Hogarth universes, machines using the expansion of “mixmaster” universes and others. Despite all these proposals, it is probably fair to say that the defenders of hypercomputing have not themselves proposed a notion of computing, they have restricted themselves to a rejection of the notion of computing expressed in the strong Church-Turing thesis.[176]

---

174 Note that it is strictly speaking misleading to talk about the computing of a “Turing machine” in this context. A Turing machine is a theoretical device that can perform a particular algorithm and the theoretical universal Turing machine is a machine that can perform whatever any particular Turing machine can perform, i.e. it can be programmed to perform any algorithm. The Church-Turing thesis concerns the possibilities of this universal Turing machine and its relation to the notion of “effective computability”. However this machine is just a model for what any mathematician with enough time and resources (paper and pencils - or tape and a read/write device) on his/her hand can compute. So, while the computer on my desk is a universal computer, its abilities are the same as that of the universal Turing machine (save its limited memory), but it is misleading to shorten this property to “it is a Turing machine”.

175 That is, not via a machine. I use “machine” and “mechanism” interchangeably in this paper, for lack of an adjective that differentiates the property of a mechanism (“mechanical”) from that of a machine (“machinical”).

176 Very useful surveys are in Copeland 1997 and 2002b, more critically Cotogno 2003, also Potgieter 2005 for the more mathematical literature. Special issues in Minds and Machines 12 (2002) and Theoretical Computer Science 317 (2004).

Possibility of Hypercomputing

The discussion about hypercomputing has focused on the question whether hypercomputing is possible in this world, given the physics of this world. A negative answer is sometimes called the "physical Church-Turing thesis" (Cotogno 2003) or also, "Gandy's thesis" (after Gandy 1980). There are many interesting problems with the view that such hypercomputing machines are possible in our world, given that the extant proposals involve infinity, such as infinite memory, or infinitely large machines, infinitely many steps, infinitely small parts, infinitely fast movement, infinitely fast information transfer, infinite amount of information transfer, infinitely precise measurement of quantum states, survival of infinite-energy states, infinitely expanding universes, etc.[177] However, this discussion can make no headway on the general question of whether hypercomputing is possible as long as no proposal is accepted. Even if one rejects a particular proposal, it is prudent to remain agnostic about the possibility of a more ingenious design. While that discussion is going on, one has to accept that it is important to distinguish between the truth of the strong and of the weak interpretation of the Church-Turing thesis, since one is discussing whether a particular proposal falls under the one but not under the other. In order to secure the traditional reading of Church-Turing, one would have to show that hypercomputing is impossible in this world, preferably in any possible world. Some attempts in this direction have been made (esp. Cotogno 2003), using Cantor's diagonal technique, but these have been rebutted successfully (Welch 2004, Orb and Kieu 2005), in my opinion. I will make a new attempt to shift the burden of proof onto the supporters of infinite hypercomputing.

Zeno Machines: Infinite Hypercomputing

Let us investigate the notion of hypercomputing through the notion of a "Zeno machine", a concept proposed by Hermann Weyl (1927). A Zeno machine is specified in such a way that each step takes a fraction of the time of its predecessor, so if the first step takes ½ a second, for example, the times for each step could be: ½, ¼, 1/8, … This machine could make a (denumerable) infinity of computing steps in finite time, in one second. It starts at time t_0, then runs through a series of steps t_n and is done at

177 Barrow (2005), ch. 10 has a useful basic survey. For a quantum proposal, see Kieu (2002), Orb and Kieu (2005). For a relativistic proposal, Shagrir and Pitowski (2003), cf. also Potgieter (2005). For a proposal of "shrinking" Zeno-machines in a Newtonian universe, see Davies (2001).

time t_1. This machine shows that we need to distinguish “in finitely many steps” from “in finite time” in the formulation of the Church-Turing thesis.
Zeno machines are repeatedly presented by Copeland as examples of possible hypercomputers (called “accelerating Turing-machines”, 1997, 2000, 2002a), and they are the most intensely discussed proposal for digital hypercomputing (cf. Ord and Kieu 2005). Zeno machines are not Turing machines since the latter produce results only once they halt after a last step, while a Zeno machine can go through infinitely many steps - though it will be “done” in a different sense, namely *in time*.

Background: Supertasks

The logical possibility of a physical object carrying out infinitely many tasks (e.g. computing steps) in finite time was much discussed in the 1950ies and 60ies on the background of Zeno’s paradoxes of movement (esp. Achilles and the tortoise, and the racetrack) and such tasks were dubbed “super-tasks” by James Thomson (1954). In order to show that performing supertasks is impossible, Thomson had proposed to consider a lamp that is switched on and off infinitely many times. He then said that from the assumption that each time the lamp is switched on it is also switched off afterwards, it follows that it can be neither on nor off after the switchings are over - which he claimed to be a contradiction. Paul Benacerraf (1962, 779ff) criticized this move, pointing out that, given the specification, nothing follows from the states of the lamp inside the series about the state of the lamp after the series.
The logical gap between what is the case inside the infinite series and what is the case after the series is crucial for the discussion and I shall call it the “Benacerraf gap”. It appears that the defender of infinite hypercomputing has to bridge the Benacerraf gap, in order to generate an output.
It is crucial for the understanding of the Benacerraf gap to keep in mind that there is no such thing as “the last step” or “the last state” in the series, and accordingly, no last step can determine the state of the lamp. Also, for any point in time arbitrarily close to time t_1, there is still a further step to take place later. Given that there is no last state, one cannot measure/read out the last state and one can not write a program that instructs “do the last step and then do this and halt”, neither can we ask “what is the state after the last step?” Since Copeland realizes this (2002a, 287 etc.), a first form of the fundamental problem is that we cannot have a computational output after the “last step” - but neither can we look at the output after the series is over in time, since “nothing follows”, as Benacerraf had pointed out. So, whatever the state of the Zeno machine at t_1, how can it be the effect of the infinite t-series, how can it be a computational state? Could we not make sure that there is an

output that can be generated without reliance on contradictory notions like “the last step in an infinite series”?

A Proposal for Infinite Computing: Facing the Benacerraf Gap

One might, for example, want to know the answer to Brouwer’s classic question (discussed by Wittgenstein) whether there is a sequence of “777” somewhere in the infinite expansion of π. This problem cannot be effectively computed because a negative answer would require looking at all of the infinite expansion of π, one by one. However, a positive answer is possible if one comes across the sequence “777” somewhere in π (we now know that 777 does indeed occur in that expansion). Many famous mathematical problems have this “semidecidable” feature, e.g. Hilbert’s Tenth Problem (claimed to be solvable by probabilistic quantum computing in Kieu 2002 and 2004) and Turing’s halting problem. Since the halting problem is precisely the problem whether the Turing machine will halt on a given problem, the *Entscheidungsproblem* itself is one of these problems.

A semidecidable task may appear to be computable by Zeno machine: Our hypercomputer may be fitted with a lamp and, for example, programmed in such a way that it switches on the lamp as soon as it finds the sequence “777” in π. After the series of computing is over, at t_1 or later, you look at the lamp: if it is on, there is such a sequence, otherwise there is not. In this fashion, any Boolean (true/false) decision over infinite domains could be settled. (And, given the possibility of binary encoding, it would appear that *any* formal problem whatsoever could be settled.)

Recall, however, that nothing followed from the specification of Thomson’s lamp about the state of his lamp at t_1 or later. Is this any different with our new, separate, indicator lamp? What the specification does tell me is that I can check whether the lamp is on at any time in the t-series, arbitrarily close to t_1: if the lamp is on, a “777” has been found. But *this* task, namely whether the sequence is to be found in π *up to a specific point*, is a Turing-computable task. Does the specification of our machine tell me what is the case with my lamp at t_1 or later? No, it does not. We have no reason to take the state of such a lamp as the output of the machine. More work needs to be done.

Copeland actually uses what I call the Benacerraf gap to avoid problematic output, saying “The answer to the Thomsonian question ‘Where is the scanner [of the Turing machine] at that point?’ is: Nowhere.” (2002a, 289). What we are told is that this machine was computing, but that we can not have an output, necessarily! A machine that necessarily has no output hardly qualifies as a computing machine: it is *hypocomputing* rather than hypercomputing.

If one wanted to provide a specification that bridges the gap, one would have to avoid *any* reference to “the last step”, and instead talk about what is the case “after the series is over in time”. I propose that we might include in the specification that there is an *indicator* (like the “lamp” above) separate from the actual machine, and we add a *bridging principle* to the effect that “the indicator is wholly determined by the machine”, in particular, it does not change other than by action of our machine. We can then check the indicator (a variable to read out, a lamp, or a display) after t_1 and use this indicator for the output of computing results.[178] Copeland and Benacerraf grant that it is logically possible to look at Thomson’s lamp after the supertask, so why not look at our separate indicator? As Benacerraf says, “Certainly, the lamp must be on or off at t_1 (provided it has not gone up in a metaphysical puff of smoke in the interval), but nothing we are told implies what it is to be.” (1962, 768).

A Second Proposal: Bridging the Benacerraf Gap

So, the bridged indicator might get us across the Benacerraf gap, but do we really want to go there? Thomson had also proposed a machine that prints the values of π on a tape that is generated at the same speed as the computation. After the end of the computing series, we would have an infinitely long tape with each digit of π printed on it. He additionally proposes a parity machine connected to the π machine, and asks “what appears on the dial after the first machine has run through all the integers in the expansion of π?” - pointing out, of course, that any output is contradictory (1954, 5). Copeland concedes that this combination with a parity machine is logically impossible, and adds: “Nevertheless, Thomson’s query as to what state an infinity machine may consistently be supposed to be in *after* it completes its supertask is a good one.” (Copeland 2002a, 286f.). So, would bridging the gap not have the unacceptable consequences Thomson wanted to warn us about? We would now be able to compute impossible things like the highest natural number, the last digit of π, the result of “0+1-1…”, etc. Let us take a closer look at the π-machine:

1) There is no last digit of π (Assumption)
2) A Zeno machine will compute an infinite number of steps in finite time (Assumption)
3) There is a program (P) such that:
   a) it calculates the digits of π one by one, and
   b) it writes each calculated digit into a variable (N), and

178 This does the job of what Earman and Norton call the “persistence property” (1996, 238ff) of the natural world, the property of persisting unchanged after the t-series. This is what causes the apparent contradiction in Thomson’s lamp, on their analysis.

c) (N) changes if and only if (P) changes it (Assumption)

4) A Zeno machine can run (P) (Assumption)

5) After carrying out (P) on the Zeno machine, variable (N) holds one digit of π (from 3 and 4)

6) That digit in (N) is the last digit of π (from 2 and 5)

7) There is a last digit of π (from 6)

Lines 1) and 7) form a contradiction. I propose the assumption we should drop is the one in line 4).

Notes: If you want to drop any part of the assumptions 2), 3), or 4), you thereby remove the specification of the bridged Zeno machine and are back at the original problem of impossible machines. If you think the bridging principle in 3c) is insufficient, then you must strengthen it, otherwise you are back to the Benacerraf gap. Equally, if you think 6) does not follow, this opens the Benacerraf gap, for any output. If you are worried about 5) being internal, you may add a line to the program where a further variable (M) is set to the value of (N) and then read out (M) after the series. Finally, note the generality of the problem since any Zeno machine that counts its own infinite steps would have to calculate “the highest natural number”.

So, while Copeland could say “No inconsistency in the notion of a π-machine was ever demonstrated” (2002a, 284), we now have a dilemma of computing with no output or bridged computing with a contradiction.

If it were the case that nothing in the argument hinges on *how* this feat of “completing” infinite steps is achieved (whether it is fractions of time, quantum superpositions, relativistic space-times, or whatever), then the argument applies to all hypercomputers that compute infinitely many steps in finite time (for an observer) – which covers all proposed digital hypercomputers.

Ways out of the Dilemma: Bridged Supertasks in Mechanisms

It may be thought that this horn of the hypercomputing-dilemma must show too much: could an infinite omniscient God not know mathematical facts over the infinite, are there no functions with truth values over the infinite? Yes, there are; what we have looked at, however, is an idealised *digital computer* with an *output* - both of these are required to produce the contradiction. We are not looking at what is computable in a some other abstract sense.

Nothing prevents God from digitally going through the extension of π in a minute. In particular, if one removes the condition of the computational output (like Shagrir 2004, 110f), no contradiction ensues. But what even God cannot do is to perform an infinite *digital* computation, say, hold up a particular finger (the original digit “indicator”) each time he computes a

digit of π *and* claim that nothing else changes the state of his fingers after the computation (a bridging principle). What would his fingers show once God is done? An infinite computing machine with a bridging principle is impossible – at least if that machine writes each digit of π, or keeps counters, or switches an indicator lamp on/off after every +1/-1 computation, etc.

But why the impossibility? Which assumption in the above argument is it that leads to the contradiction and needs to be abandoned? The burden of proof is now on the defenders of hypercomputing, who have to explain the specification of the machine such that it does have an output but does not result in contradiction.

Finding a way out is not as easy as it may seem. It is not enough just to *state* that a machine, say, computes the "777" problem, and to *state* that the indicator displays the solution at t_1 and after. The problem is *not* just one of mathematical possibility. The possibility of hypercomputing involves more than a formal specification of the algorithm that is free from contradiction; it involves the possibility *as a digital mechanism*, i.e. as a mechanism in which the state of the indicator at t_1 and after is *causally determined* by the step-by-step workings of the mechanism. Put in these causal terms, a bridged supertask is one where the supertask has an effect that lasts beyond the time of the completion of the task, an effect that can be taken as the output of the computation.

So, if we want to uphold the possibility of supertask hypercomputing, we must find a specific reason why there can be a Zeno-machine that cannot run the programs resulting in contradiction, but can run others. To be sure, for a machine that deals with semidecidable problems the results themselves cannot be a source of contradiction (no single answer to a well-formed yes/no question results in a contradiction). The only difference between the manifestly impossible machines and the proposed machines for semidecidable tasks appears to be that the bridged indicator is changed infinitely many times in the former and only once (if at all) in the latter - though perhaps infinitely close to t_1. In fact, the indicator itself is performing a supertask in the impossible machines! Is there an explanation that would allow for the semidecidable machines, but rule out the manifestly impossible machines? Let us have a brief look at some candidates:

a) Nothing can be the effect of infinitely many causes[179]

This would prevent the existence of a bridged indicator for the impossible machines that has been updated infinitely many times. It also implies that the Zeno-machine must go up in Bencerraf's "puff of metaphysical smoke" after the t-series, since its state at t_1 would be a result of infinitely many steps. But it would also show that the state of the indicator cannot

179 Meaning "in finite time". There can be no effect after an infinite time (for an observer) anyway. (But see Hamkins and Lewis 2000 for an investigation of what is mathematically possible.)

be the effect of each and every one of the infinite steps of the Zeno-machine, therefore infinite hypercomputing is impossible.[180] In fact, it implies a rejection of assumption 4) in 2.3 above, so it is not a way out.

b) Nothing can have infinitely many effects

This would prevent the Zeno-machine from updating the indicator infinitely many times. But it would also prevent any Zeno-machine from running, since starting it would have infinitely many effects. Caused supertasks and thus infinite hypercomputing would be impossible. This also implies a rejection of 4), so it is not a way out either.

c) If something changes infinitely many times, then it must go out of existence “afterwards”, without any effect

This is a consequence of a), but strong enough to prevent a bridged indicator that has been updated infinitely many times, while allowing a bridged indicator that is changed once. It would force the Zeno-machine itself to cease existence after its activity (as is the case in the non-Newtonian proposals). It requires that performing a supertask is possible, but makes an object cease to exist *without any effect*. The going out of existence cannot be the consequence of internally “having completed” the supertask, but neither can it be caused from the outside by “time is up”. On the other hand, if going out of existence is due to some gradual process, it becomes physically implausible that an object should have effects infinitely close to its going out of existence. - I think that explanation c) makes too many ad hoc assumptions, but we should be open to arguments sustaining it. As long as these are not forthcoming, I propose to reject c).

To sum up, bridged supertasks require both that something can be the effect of infinitely many causes (~a) and that something can have infinitely many effects (~b). Denying one of the two conditions (a, b) amounts to denying the possibility of bridged supertasks, thus the denial of 4) in the argument above - for any program (P). I do not see any further plausible explanation. If there really is no other, we have to conclude that bridged supertasks are impossible - whether or not they are considered computing machines. To put it the other way ’round, if bridged supertasks were possible, infinite hypercomputing should have been possible. But infinite hypercomputing is not possible, so bridged supertasks must be impossible.

180 The specification of the mechanism must be such that the state of the indicator lamp at t1 can be taken as the “output” of the computing procedure. This applies when the lamp is “off” as well as when it is “on”. Not changing the indicator after a particular computation is also an effect. We cannot take the indicator as output if the causal connection was cut somewhere during the t-series - and this cut caused the lamp to be “off”. This would return to the other horn of the dilemma: computing without an output.

Conclusion: Computing the Incomputable and Other Supertasks

If the general conclusion could be established that bridged supertasks are impossible (by a thorough rejection of all ways out), this would have ramifications for Zeno's classic paradoxes. If principle a) is true, it cannot be the case that moving through a stretch from a point A to a point B is to perform a supertask, since the arrival at B presumably is the effect of moving through that stretch - and we do not vanish as soon as we complete a movement from A to B. Equally, if we prefer only principle b), and take moving through a stretch from a point A to a point B as a supertask, then our movement cannot have a cause - but it obviously can have a cause. If c) is true, after all, we would have to vanish after a movement. So, on *any* of these three explanations Zeno must be wrong when he says that one movement is to make infinitely many movements.

Concerning our original question, I conclude that Thomson was right that if (!) anything follows from states inside the series to states outside the series, then contradiction ensues. And Benacerraf was right that nothing does follow from states inside the series to states outside the series - unless one adds a bridging principle. In other words, either the Zeno machine can be specified, does bridge the gap, but then its specification involves contradictions, or it is underspecified, does not bridge the gap, but then it does not compute an output. Either way, Zeno machine hypercomputers are impossible - as are probably all bridged supertasks. Therefore, the notion of infinite hypercomputing is no reason to reject the traditional strong interpretation of the Church-Turing thesis.

Bibliography

Barrow, John D. (2005) The Infinite Book: A Short Guide to the Boundless, Timeless and Endless. New York: Pantheon Books.

Benacerraf, Paul (1962) "Tasks, Supertasks, and the Modern Eleatics". Journal of Philosophy LIX/24: 765-784.

Boolos, George S., John P. Burgess and Richard C. Jeffrey (2003) Computability and Logic, 4th corr. ed. Cambridge: Cambridge University Press.

Bringsjord, Selmer & Zenzen, M. (2003) Superminds: People Harness Hypercomputation, and More. Dordrecht: Kluwer.

Churchland, Paul (2005) "Functionalism at Forty: A Critical Retrospective". Journal of Philosophy 102/1: 33-50.

Copeland, David (1993) Artificial Intelligence: A Philosophical Introduction. Oxford: Blackwell.

— (1997) "The Broad Conception of Computation". American Behavioral Scientist 40: 690-716.

— (1998) "Turing's O-Machines, Penrose, Searle and the Brain". Analysis 58: 128-138.

— (2000) "Narrow Versus Wide Mechanism, Including a Re-Examination of Turing's Views on the Mind–Machine Issue". Journal of Philosophy 97/1: 5–32. (Reprinted in Scheutz 2002.)

— (2002a) "Accelerating Turing Machines". Minds and Machines 12: 281–301.

— (2002b) "Hypercomputation". Minds and Machines 12: 461–502.

— (2004) "Computation". In: Luciano Floridi (ed.): Blackwell Companion to the Philosophy of Computing. Oxford: Blackwell, 3-17.

Copeland, David and Diane Proudfoot (1999) "Review of: 'The Legacy of Alan Turing', ed. Peter Millican/Andy Clark". Mind 108: 187-195.

— (2000) “What Turing Did after He Invented the Universal Turing Machine”. Journal of Logic, Language and Information 9: 491-509.
Cotogno, Paolo (2003) “Hypercomputation and the Physical Church-Turing Thesis”. British Journal for the Philosophy of Science 54/2: 181-223.
Davies, E. B. (2001) “Building Infinite Machines”. British Journal for the Philosophy of Science 52/4: 671-682.
Davis, Martin (2000) The Universal Computer: The Road from Leibniz to Turing. New York: W. W. Norton.
Fodor, Jerry: (2000) The Mind Doesn’t Work that Way: The Scope and Limits of Computational Psychology. Cambridge, Mass.: MIT Press.
— (2005) “Reply to Steven Pinker ‘So How Does The Mind Work?’” Mind & Language 20/1: 25-32.
Gandy, Robin (1980) “Church’s Thesis and Principles of Mechanics”. In J. Barwise, H. J. Keisler and K. Kunen (eds.): The Kleene Symposium. Amsterdam: North Holland.
Hamkins, Joel David and Andy Lewis (2000) “Infinite Time Turing Machines”. The Journal of Symbolic Logic 65/2, 567-604.
Harel, David (2000) Computers Ltd.: What they Really Can't Do. Oxford: Oxford University Press.
Kieu, Tien D. (2002) “Quantum Hypercomputability”. Minds and Machines 12: 541-561.
— (2004) “Hypercomputation with Quantum Adiabatic Processes”. Theoretical Computer Science 317, 93-104.
Ord, Toby and Tien D. Kieu (2005) “The Diagonal Method and Hypercomputation”. British Journal for the Philosophy of Science 56/1: 147-156.
Piccinini, Gualtiero (2004) “The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts’s Logical Calculus of Ideas Immanent in Nervous Activity”. Synthese 141/2: 175-215.
Pinker, Steven (2005) “So How Does the Mind Work?” and “A Reply to Jerry Fodor on How the Mind Works”. Mind & Language 20/1: 1-24, 33.
Potgieter, Petrus H. (2005) “Zeno Machines and Hypercomputation”. Online 21.09.2005, http://arxiv.org/abs/cs/0412022
Scheutz, Matthias (2002) (ed.) Computationalism: New Directions. Cambridge, Mass.: MIT Press.
Shagrir, Oron (2004) “Supertasks, Accelerating Turing Machines and Uncomputability”. Theoretical Computer Science 317: 105– 114.
Shagrir, Oron and Itamar Pitowsky (2003): “Physical Hypercomputation and the Church–Turing Thesis”. Minds and Machines 13/1: 87-101.
Siegelmann H. T. (1995) “Computation Beyond the Turing Limit”. Science (268), 28 April, 545-548.
Siegelmann H. T. and Sontag E. D. (1994) “Analog Computation via Neural Nets”, Theoretical Computer Science 131, 331-360.
— (1998) Neural Networks and Analog Computation: Beyond the Turing Limit. Boston, Mass.: Birkhäuser.
Thomson, James F. (1954) “Tasks and Super-tasks”. Analysis 15: 1-13.
Turing, Alan (1936/1937) “On Computable Numbers, with an Application to the Entscheidungsproblem”. Proceedings of the London Mathematical Society 2/42: 230-256.
— (1992) Collected Works: Mechanical Intelligence. Amsterdam: North-Holland.
Welch P.D. (2004) “On the Possibility, or Otherwise, of Hypercomputation”. British Journal for the Philosophy of Science 55, 739-746.
Weyl, Hermann (1927) Philosophie der Mathematik und Naturwissenschaft. München: R. Oldenbourg. [English translation, Princeton UP, 1946.]

THE DOCILE HACKER

THE OPEN SOURCE MODEL AS A WAY OF CREATING KNOWLEDGE

EMANUELE BARDONE[a], LORENZO MAGNANI[ab], AND DAVIDE SECCHI[c]

[a]*Department of Philosophy and Computational Philosophy Laboratory, University of Pavia, Italy, bardone@unipv.it* [b]*Department of Philosophy, Sun Yat-Sen University, Guangzhou, P.R. China, lmagnani@unipv.it* [c]*Department of Economics, University of Insubria, Varese, Italy, dsecchi@eco.uninsubria.it*

Introduction

Historically, what it is commonly labeled as *Open Source* (OS) embraces an astonishing variety of methods that cannot be reduced to one single approach. The great number of licenses under which work can be released is just one example demonstrating this fact. *Free* and *Open Software*, *GNU* and *GPL* licenses, *Creative Commons*, *Copyleft*, *Open Standards*, are just some of the different projects that belong to the *Open Source* galaxy. In this paper we do not aim at examining all the differences between these approaches and ideas. Instead, we simply refer to the term "Open Source (OS) Model" as a general mode of knowledge transmission and creation that is based on one very simple idea: the *source code* of a software must be *visible* and editable so that it can be used, redistributed, changed, and upgraded by everybody.[181]
What strikes us most about the various OS projects is the tremendous success they have gained during the last two decades. *Linux* is indeed

181 As computer scientists know well, it is common to distinguish between the so called source code and the object code. The former is the human-readable, human-editable form of a program. The latter is what the computer actually runs, and therefore what humans interact with (Raymond, 2004).

the most well known case, but there is a number of other OS projects from web servers (such as APACHE) and office suites (such as OpenOffice) to script languages (such as PHP), from databases (such as MYSQL and POSTRESQL) to protocols (such as TCP/IP), that have now become leading products within their sector. Although Bill Gates forecasted that the "hobbyist" of the computer (as hackers were called by Gates) would have soon disappeared, the success of OS software and products is now widely recognized as a major event in the history of computing and business.[182]

Now, the point we want to develop in this paper is whether "being open source" might derive from the way individuals process information, and relate to each other. What is the cognitive basis of individuals operating "open source"? Is there any social cognitive motivation for "being open source"?

In this paper we try to provide what seems to be a useful cognitive background for the individuals, where we attach what defines their social behavior. This is a first step towards a detailed explanation of the open source as a plain externalization of individuals' cognitive processes. In order to show our approach, we divided the paper into two parts. Even if we think that the two parts integrate in a fruitful way, they remain quite different. The first section of the paper is dedicated to the distributed cognition approach, and to what we define as its behavioral counterpart (i.e. the "docility" attitude). The second section relates to the OS and to the way it can be defined in terms of docility. Finally, while in the first part we do not need to mention OS so much, in the second we do.

The main assumptions regarding the human cognitive system

Our starting point has been that of focusing on the way the human cognitive system works in a social setting. This interest stems out from the idea that the successful diffusion and the workability of the open source model – it is a model through which knowledge is organized – can be traced through the way our cognitive system really works.

This section is dedicated to the analysis of a new way of thinking about and modeling the cognitive system. This implies some "social attitudes" that we describe as "docile attitudes" of the individuals (Simon, 1990; 1993).

182 This definition can only regard programmers and people involved in computing. However, the principle that lies behind Open Source can also be generalized and applied to other domains, as Wark (2004) has shown brilliantly.

Beyond the "cognitive divide"

The human cognitive system has undergone much remodeling in the past fifty years. Numerous metaphors have been provided (such as that of the computer), and many approaches have been developed.

There are three main classes of theories that Richardson (2000) classifies as (1) computational, (2) connectionist, and (3) social cognition theories. It is not the aim of this paper to analyze these approaches, however. One of the points here is that all these theories are based on the assumption that there is a divide between *internal* and *external* resources to the human cognitive system. We claim that this divide is probably useful for didactical purposes, but misleading in terms of explaining how the cognitive system really works.

Following Clark and Chalmers' (1998) intuition, we argue that "any cognitive activity cannot be regarded only as an internal process that occurs within the isolated brain." This statement leads to two basic points. The first relates to human cognitive capabilities; these are not limited to the brain. The way we interact with external objects (artifacts, thoughts, or any other external element) shapes our cognitive system. So that what humans have inside their body is not *de facto* divided by external resources. Cognitive status changes together with external stimulations; moreover, it is organized with or shaped by external resources. Broadly speaking, we overcome our limitations – that we can term bounds, according to Simon (1955; 1979) – through the exploitation of external resources. Humans have computational and cognitive limits, but they depend on the use of external resources. From this angle, it is not enough to state that the cognitive processes are not merely internal facts, but we argue that they cannot even exist without external resources. Everything happening inside the human brain relates to external facts, artifacts, thoughts, speeches, writings, things, and so on, i.e. it relates to external resources (Knuutilla, and Honkela, 2006).

The second issue relates to the role of these resources. We can, for example, classify these resources relating to their (a) proximity to the process, (b) material or intangible nature, or (c) to their social or individual character. Whatever their nature, the most important element is that cognitive processes also occur outside the brain. This means that external objects acquire a cognitive role; they operate as part of the system, but are located outside the human brain. We may say that they are a kind of cognitive extension that helps humans to overcome their cognitive limits (Magnani, 2006a; 2006b). For example, part of these resources play the role of memory storage, so that individuals do not have to keep concepts in mind but rely on external supports, on occasion.

The cognitive system is then *distributed*, in the sense that many cognitive functions are located in external resources.[183] Thus, human cognition leans on external resources; this means that there is an enhancement of human potential, depending on their exploitation (Hutchins, 1995). How do we overcome everyday cognitive difficulties? The answer is very easy: we exploit external resources. For example, can a non-skilled individual solve a logarithm without any external support? They can, but it would be easier for them to write down the process, or to use a calculator. And, in this case, what do they do do? They use external resources in order to overcome their computational limitations. This simple example shows that individuals ordinarily behave this way. Moreover, we may add that individual limits cannot be reduced to computational problems.

What the example shows is the *distribution process*: we distribute cognitive capabilities to external supports or resources (Magnani, 2006b). If this is the way the cognitive system operates, i.e. if we are used to basing our cognitive activity on external resources, it follows that we somehow create part of these external resources. The act of creating external supports is called the *externalization process*. We go back to the logarithm example: what is the cognitive meaning of writing down a logarithm on a sheet of paper? When writing, we put something internal outside of us, i.e. we externalize. Hence, the fact that we lean on external resources leads us to create part of these resources. In summary, we can state that (1) we lean on *existing* external resources, and (2) we *create* external resources (Hutchins, 1995; Magnani, 2001; 2006b).

Cognitive distribution is based on a strict interaction between the internal and the external resources, where the whole cognitive system is shaped by external resources. This leads to the essential fact that individual behavior and thinking need external resources; where they do not exist, individuals create them.

Let's take a look at the other side of the coin. Once we externalize, or we have exploited an external resource, we tend to re-internalize. The draft, sentence, painting, idea, etc. that the individual first externalized, change their meanings, once externalized. What then? What we have externalized (the logarithm) becomes something different; it is something objective, different from its original form (when it was inside the brain).[184] Once externalized, we play with these external resources, and then solve, for example, the logarithm. This process is that of *re-projecting* internally what has occurred outside, in the external invented structure (Magnani, 2006a; 2006b). This is the process that leads to new ways of thinking, e.g. the solution of the problem we have with the logarithm.

183 The fact that we use the terms "cognitive system" instead of "brain" or "mind," follows the fact that we do not intend to physically define human intellectual activity.

184 Moreover, it becomes understandable, readable, intelligible, etc.; in other terms, it becomes a social resource (tool).

In summary, we have divided the distributed cognition approach into two main processes: (a) externalization, and (b) re-projecting. We claim that this approach helps in defining the way the human cognitive system works, and also many variables of social interaction.

The docile cognitive system

The theory of distributed cognition is relevant for the understanding of individual cognitive processes. In this work, we argue that it reveals a potential to explain the social side of human behavior too. In this way, what are the kinds of human behavior that best fit the way our cognitive system actually works? The answer does not appear to be so straight forward.

Our hypothesis is that the cognitive system is distributed, and this fact leads to the exploitation of the external resources that we find in the environment (social, natural, and so on). If we lean on existing external resources, then we also create some of them, through the externalization process. Once externalization is accomplished, we then re-project information with a re-internalization of data. Thus, on the one hand, human behavior needs to support these processes and, above all, enhance externalizations, whatever form they might take. On the other hand, a significant part of external resources are connected to other individuals, and much individual externalization has a high probability of being "socially oriented" (i.e. it has to do with the social environment). In short, human behavior supports the way our cognitive system works through (a) its orientation towards externalizations, and (b) being socially oriented (Bardone, and Secchi, 2006; Secchi, 2005). Let's focus on these two basic points.

If we think about the way the externalization process occurs, we may think of an individual trying to solve a problem. When we are looking for a new house to live in, for example, there is at least one step in which we need to externalize. When we read the newspaper, we draw red circles to isolate the most interesting properties. This simple action changes the external resource, and the meaning we confer to the circled property. In other terms, this is the way the learning process evolves. The selected circled property serves as a *cognitive mediator*, and it is an external resource, as we modified it.

Many cognitive mediators preserve a similar meaning for different individuals (Hutchins, 1995). This fact constitutes the basis of social interaction, since human beings lean on external resources that have a social meaning. Cognitive mediators are social and external resources when they connect two or more individuals on the basis of the mediation process. This is a common sense statement. Think of a pen: why do we know how to use it properly? It's easy: because we have seen someone using a similar pen or pencil before. The exploitation of this external

resource depends on a social event that occurred in the past. This is typical of human behavior, and it is easier to lean on something learned than to attain this learning by ourselves.

What we call "docility"

The way we learn dramatically depends on the social environment. The cognitive basis of our understanding draws on the exploitation of social channels (and cognitive mediators). So that, on the one hand, individuals lean on information, recommendations, and suggestions that are socially transferred (passive; Simon, 1993), and they transfer information, recommendations, and suggestions through social channels (active; Bardone, and Secchi, 2006). We have an active and passive way of interacting with other individuals. What emerges is that this type of interaction seems to be strictly related to the way we described the processes underlying the human cognitive system.

We can, now, rewrite the question with which we started this section: what kinds of behavior best fit the way our cognitive system actually works? We use the word *docility* to describe all the active and passive attitudes that individuals demonstrate towards the learning processes and their everyday courses of action (Simon, 1990; 1993). As we define it, docility is the tendency to depend on suggestions, perceptions, comments, and to gather information from other individuals (or through social channels, that is a wider approach to the issue), on the one hand, and to "provide" them, on the other (Bardone, and Secchi, 2006).

This tendency, i.e. being docile, supports the way the human cognitive system works. How can we define "suggestions, perceptions, and comments"? They are nothing but external resources; and they are of a definite kind, because they belong to other individuals. In other words, being docile (on the passive side, i.e. other individuals providing suggestions and so on) means to lean on these external resources, so that our cognitive status changes together with their exploitation. On the active side (i.e. when we provide suggestions, comments, etc.), we are externalizing something. It is worth noting that the two processes are not disconnected, in the way we ordinarily behave. We interact with other individuals and, since language exists as an externalization, we cannot divide, in practice, externalization from the exploitation of external resources and re-projecting. On the contrary, we do it for explanatory reasons.

Interestingly, docility also focuses on the type of the resource we refer to. Being docile means to base a proper action on the exploitation of social channels. These channels have been widely used, since the beginning of human life. However, we do not want to focus here on the nature of these channels, but on their role as external resources. The main fact here is

that external supports (artifacts, tools, etc.) tend to pass their docility to the (social and cognitive) system they belong to, increasing the cognitive chances available in it. This tendency is called the *docility effect*. In other terms, we argue that there is an "organizational" way that allows docility (and our cognitive system) to work as a major attitude of individuals. As many studies on this issue try to underline, cooperative, altruistic (Knudsen, 2003; Simon, 1993; Khalil, 2004; Secchi, 2005), and socially responsible behavior (Secchi, forthcoming) do derive from the docile attitude of individuals.

Docility tends to become structured in organizations of various kinds (State, firms, associations, and so on). This means that organizations maintain some sort of mechanisms that foster individual docile attitudes. Every organization, for example, can be defined in terms of the social channels that group data, the top-down relations, behavioral or normative codes of conduct, the way people tend to cooperate or not, and so on.

Our main hypothesis is that organizations that embed more than one docile "mechanism" tend to have a better fitness, both in terms of the ways individuals exploit these mechanisms, and in relation to the social environment. This is the way the open source works, and is the reason for its success.

The open source model and its (cognitive) kernel

A matter of cognitive reliability: why a cognitive account is needed

During the last few years economists and sociologists have provided stimulating accounts that try to explain the success of the OS movement and its rationale. Raymond (2001), for instance, pointed out that the radical innovation of the OS model was social rather than merely technical. To explain this idea, he introduced an illuminating metaphor that clearly depicts the culture and the values of the OS model. Most of the companies involved in programming – he argued – resemble what he called "a reverent cathedral building" with a rigid hierarchy. In contrast, the OS model is more like a "babbling bazaar of different agendas and approaches". No rigid hierarchies, no bosses, but very committed users that report bugs, and are also able to fix them and suggest alternative solutions or new problems to solve. On another note, Himanen (2001) provided a sociological account in which he compared so-called hacker ethics with Protestant work ethics and drew some interesting conclusions about the impact that this new radical approach may have on existing theories of business.

Although that sounds most appealing, these kinds of accounts do not hold water, because they fail to put forward any explanation about why "being open source" can also be extremely successful, from a cognitive perspective. For the main task is to investigate the cognitive reliability of the OS model and open up its *cognitive kernel*. Generally speaking, the main idea is that *being open source* may be something more than a business philosophy or a type of work ethic: it may also match a general trait of human cognition in the way it works and evolves. The point we want to make is the need for a cognitive account of the success of the OS model. Generally speaking, the OS movement deals with information and knowledge transmission: therefore the way it manages, organizes, and extends cognitive abilities to cope with programming becomes a crucial aspect that cannot be neglected.

In the following section, we describe the four dimensions in which the notion of docility may be a valuable candidate in explaining the cognitive relevance of the OS model.

The docile Hacker

Sharing code

The very idea of the OS model is that the source code of a software must be visible and editable so that it can be used, redistributed, changed, and upgraded by everybody. Now, the point we want to make in this subsection is that sharing code is a product of docility.

Source code is not just a block of bits that saves time for those who, fortunately, can use it. Source code is a *cognitive repository* that stores ideas, problems, trials as well as errors, solutions, and it may suggest alternative views. If that is correct, then *sharing code* contributes to releasing a large body of knowledge and information that drastically modifies how other people (in this case, hackers) can *learn*, *solve* problems, and more generally accomplish a cognitive task such as that of making up computer programs. As Raymond put it "you often don't really understand the problem until after the first time you implement a solution" (2001: p. 25). In doing this, hackers lean on various external resources (in this case, the source code written by others) that become a major basis for their cognitive work and performances. That is exactly what docility is all about. That is, in writing and then sharing the code, hackers are continuously involved in a "smart interplay" between their brain and the environment that is facilitated and enhanced by a tendency toward external resources: that is docility.

Indeed, in the case of proprietary software, programmers share code and, to some extent, they are docile as well, because they take

advantage of others' improvements. However, docility is limited by the narrow boundaries of the company they work for: nobody else can access the code. In contrast, hackers can potentially rely on thousands of people all committed to the same problem[185]. An example of this enormous potentiality is given by the high reliability that open software guarantees to the user. As Raymond wrote "many eyeballs tame complexity" (Raymond, 2001 and 2004). As a matter of fact, Microsoft products (from computer servers to PCs) are much less reliable than Linux in terms of security, scalability, performance, compatibility, stability, and so on.[186] In proprietary software companies, the fact that docility is limited jeopardizes all the cognitive benefits provided by docile behaviors. For instance, peer review, that is indeed one of the most successful factors leading to software reliability, is dramatically reduced. In fact, the peer review principle is based on the possibility for everybody to check each other's work without limitations of any sort. None can hide his/her work and prevent others from criticizing it. In this sense, secrecy is the enemy of quality and it can be regarded as highly anti-docile behavior.

Building communities

We argue that code-sharing contributes to releasing a great portion of knowledge that drastically shapes the cognitive task hackers face. The same can be said for another feature of the OS model, not in this case connected to inanimate resources (the code) but animate ones, that is, other human beings. In this case, docility is crucial to making use of those cognitive resources embedded in social channels. That is, hackers are docile in the sense that they do not simply work on the same piece of code: they build up communities of practice in which learning from others and then teaching what is experienced becomes a major trait in the way knowledge is transferred and developed. In this sense, the social dimension turns out to be a significant cognitive source for their work. Peer review is indeed an example of this kind, as briefly discussed above: people that get involved in an open project release their work openly to other hackers that in turn provide them with suggestions or improvements or simply test their distribution. Cooperation is therefore a direct consequence of the way they work, not only an "ethical" option (Himanen, 2001: p. 68).

---

185 In a famous paper of his, Bill Gates (1976) argued that hobbyists could not have built up reliable and stable computer programs. They must have been paid for doing such a good job. On this note, he wrote: "Without good software and an owner who understands programming, a hobby computer is wasted. Will quality software be written for the hobby market?"

186 See, for instance: http://www.hesketh.com/publications/a_winning_argument_for_linux.html

Docility is also displayed in the ubiquitous use of social tools such as forums, chat rooms, mailing lists, newsgroups, newsletters, etc. As a matter of fact, for any open source project there is a community of practice and learning. Hackers and developers are allowed to exchange information, solutions, suggestions, know how, etc. As a matter of fact, most of the activities concerning software development are managed and organized through the Internet. Usually, open source projects start out from a person or a group of people that stumble over a series of unsolved problems. Then, they post some information about their problems on a website or a mailing list and try to get some help from other hackers. This gives rise to a community in which hackers can freely cooperate on the project or simply get an idea of what is going on. Thus it is not surprising that historically the success of the OS Movement was largely due to the creation and implementation of tools that enabled distance communication.

In this sense, forums, chat rooms, and the like are cognitive mediators that encode and then release a great portion of resources embedded in social channels and facilitate knowledge transmission at the same time.

Publicly releasing new developments

As argued in section 3.2.1, source code can be considered a cognitive repository that is open to everyone who wants to modify or simply re-use it. But that is not the whole deal: the success of the OS Model is also related to the tremendous developments that it brought about. For instance, when the first version of Linux OS came out in 1991, it consisted of only 10.000 lines of code. Just after 7 years, it was made up of more than one and half million lines .[187] What does this mean? It means that the OS Model is not only about sharing code, but it is also a development model in which progress is really made possible by the thousands of hackers involved in various open source projects. That is, hackers do not only re-use and share code, but they are committed to sharing any development or contribution that may improve the quality of a software. According to the GNU General Public License, any modification made upon every single piece of code must be released, since everybody must give the recipients all the rights that are given to him/her[188]. Here again the role of docility is crucial in describing the cognitive relevance of this attitude; hackers are docile in the sense that they opt to publish their improvements for further inspections, to fix some bugs or add new features.

---

187 For more information on that, http://en.wikipedia.org/wiki/Source_lines_of_code

188 See GNU Lesser General Public License available at: http://www.gnu.org/licenses/lgpl.html

Standardizing work and the notion of standard-fidelity

Standard-fidelity is one of the most interesting aspects concerning docility. In order to introduce this notion, let us make an example. Consider the difference between a mathematical theorem and a magic trick; they simply differ in the method or the procedure they carried out in order to get results. A mathematician has to follow rules and procedures embedded in the practice that are somehow accepted as objective. For instance, one cannot use theorems that have not yet been clearly demonstrated Any passage must be justified according to the laws of logic: neither contradiction nor partiality can be accepted. In contrast to that, a magic trick is something completely private – That is to say: first of all, it is not publicly available to everybody who wants to know about it; secondly, the procedure through which one can make the trick work is kept secret as well, *known only to those within the magic circle*; third, there is no standard at all, since any magician can perform tricks on their own. Generally speaking, we may say that there is some kind of practice, such as proving a theorem or making a scientific experiment, that requires people to follow how certain resources have been employed by others and, finally, the rules embedded in them. In the case of the mathematician, he/she has to follow certain standards of mathematics, accepted within the field. In this case, docility is represented by what we call *standard-fidelity*.

The cognitive relevance of following standards or/and standardizing one's own work is as follows: first of all, using standards makes information and knowledge transmission much easier. This is true for humans and also for machines. Consider the case of formal languages comparing with informal ones. Mathematicians, logicians, etc, try to make up ways of communicating and transferring knowledge that are transparent (standardizing) as much as possible to overcome the ambivalence and ambiguity of natural languages such as English, French or Italian. Secondly, having standards also makes it much simpler to compare different claims. Consider a scientific experiment: here scientists follow certain standardized procedures that clearly display results and the way to test their presumed validity. Very often the incommensurability between theories is due to the failure to apply set standards when measuring the different claims and to then decide upon the best method. Thirdly, standards facilitate further developments. Here again, standardized procedures lead to results that can be understood more easily and shared better by the community of practitioners (scientists, mathematicians, and so on). Now, let us turn back to hackers and the relevance of standard-fidelity for the OS model.

Usually *Open Source* is viewed as something related to software development (see, for instance, section 3.2.1 and 3.2.3), whereas *standards* regard *common agreements* that allow communications

between different means (Krechmer, 2005). From an analytical perspective we do not find reasons to reject this distinction.[189] But from the hacker's point of view things starts blurring. The main motivation that stands behind the very idea of the Open Source is to keep the source code open and available to everybody for inspection and modification. Therefore, it is ultimately committed to enabling people to use and exploit all the cognitive functionalities that a software can give, without any restriction. If that is correct, then building up standards that amplify interoperability (interaction), cross-platform compatibility, usability, and so on, is a part of the OS kernel.
Now, focusing more on standard-fidelity in computing, we find two main levels at which it operates. The first one regards the kind of standard-fidelity displayed by mathematicians. As argued above, mathematicians play their game by the rules of the discipline that are not personal or subjective. The same happens in making software. Since in the Open Source galaxy a piece of code should be easily shared and modified by all, some basic requirements must be met to increase *re-usability*. These basic requirements regard, for instance, writing code (consistency and clearness, for instance). Some of them are also related to releasing pieces of code under some open source license. For instance, GNU Free Documentation License regulates verbatim copying, modifications, the documentations to release with the code, and so on.
The second aspect of standard-fidelity explicitly concerns the increasing of open standards as a major opportunity to disseminate and distribute knowledge and cognitive capabilities. There are many projects concerning open standards. Among them, it is worth citing the case of Open Document Format [190] (ODF) developed by the OASIS[191] industry consortium; ODF has been recently approved by ISO[192] (International Organization for Standard) as the first standard for editable office documents. Another well-known example is the World Wide Web Consortium,[193] primarily devoted to developing standards for the Web.

Conclusion

We presented four dimensions of the OS: (1) sharing code, (2) building up communities, (3) publicly releasing new developments, and (4) the notion of standard-fidelity. These define two points at the same time:

---

189 For an interesting debate about the distinction between Open Source and Open Standard, see Coyle (2002), Schwartz (2003), Saint-André (2003), and Corrado (2005).

190 http://en.wikipedia.org/wiki/OpenDocument

191 http://www.oasis-open.org/home/index.php

192 http://www.iso.org/iso/en/ISOOnline.frontpage

193 http://www.w3.org/

The first point relates to the fact that given these four points together, the OS movement became a model for success in software creation and, broadly speaking, knowledge management.
The second point of interest is related to the concept of docility. The OS can be usefully defined in terms of docility. The OS embeds some of the basic characteristics of what we defined as docile attitude in individuals and, in so doing, it enhances individuals' cognitive capabilities. This is also a fundamental explanation of its success.
Our conclusion becomes self-evident once these two paragraphs have been read together. Docility explains the success of the OS as a way of creating and managing knowledge. Whether docility can be used to analyze social and knowledge systems' effectiveness is the basis of our future research.

References

Bardone, E., and Secchi, D. (2005). Extending the bounded rationality model. The distributed cognition approach. *Working paper*, 9, Dept. of Economics, U. of Insubria. Available at: http://eco.uninsubria.it/dipeco/quaderni/files/QF2005_9.pdf

Bardone, E., and Secchi, D. (2006). The Distributed Cognition Approach to Rationality: Getting the Framework. Presented at the *2006 Academy of Management Annual Meeting*, Atlanta.

Bonaccorsi, A. and C. Rossi, (2003). Why Open Source software can succeed, *Research Policy*, 32: 1243—1258.

Chawner, B. (2005). Free/Open Source Software: New Opportunities, New Challenges. Available: Available at: www.vala.org.au/vala2004/2004pdfs/33Chawn.PDF.

Clark A., and Chalmers, D.J. (1998). The extended mind. *Analysis*, 58: 10-23.

Gates, B., (1976). An Open Letter to Hobbyists. Available at: http://www.blinkenlights.com/classiccmp/gateswhine.html.

Goodman, D., (2004). The Criteria for Open Access, *Serials Review* 20(4).

Graham, P., (2004). Hackers and Painters. Big Ideas from the Computer Age, O'Reilly, Sebastopol (CA).

Himanen, P., (2001). *The Hacker Ethic*. A Radical Approach to Philosophy of Business, Random House, New York.

Hutchins, E. (1995). *Cognition in the Wild*, Cambridge (Ma.): MIT Press.

Khalil E.L. (2004). "What is Altruism?" *Journal of Economic Psychology*, 25: 97-123.

Knudsen, T. (2003). Simon's selection theory: Why docility evolves to breed successful altruism. *Journal of Economic Psychology*, 24: 229-244.

Knuutilla, T., Honkela, T. (2006), Questioning external and internal representation: The case of scientific models. In L. Magnani and R. Dossena (eds.), *Computing, Philosophy and Cognition*: 209-226. London: King's College Pubblications.

Krechmer, K., (2005). Open Standards Requirements, in Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

Magnani, L. 2006a. Mimetic minds. Meaning formation through epistemic mediators and external representations. In A. Loula, R. Gudwin, and J. Queiroz (eds.), *Artificial Cognition Systems*. Idea Group Publishers, forthcoming.

Magnani, L. 2006b. *Knowledge as a Duty. Distributed Morality in a Technological World*. Cambridge: Cambridge University Press (forthcoming).

Moody, G., (2001). *Rebel Code. Linux and the Open Source Revolution*, Perseus Publishing, New York.

Raymond, E. S., (2001a). *The Cathedral and the Bazar*, O'Reilly, Sebastopol (CA), 2nd Edition.

Raymond, E. S., (2001b). *How to Become an Hacker*, available at: http://www.catb.org/~esr/faqs/hacker-howto.html.
Raymond, E. S., (2004). Open Minds, Open Source, *Analog*, June-July.
Richardson, K. (2000). *The Making of Intelligence*. New York: Columbia University Press.
Secchi, D. (2005). Altruism and selfish behavior. The docility model revisited. *Working paper*, 8, Dept. of Economics, U. of Insubria. Available at: http://eco.uninsubria.it/dipeco/quaderni/files/QF2005_8.pdf
Simon, H.A. (1955). A Behavioral Model of Rational Choice, *The Quarterly Journal of Economics*, LXIX: 99-118.
Simon, H.A. (1979). Rational decision making in business organizations. *American Economic Review*, 69: 493–513.
Simon, H.A. (1990). A mechanism for social selection and successful altruism. *Science*, 250(4988): 1665-1668.
Simon, H.A. (1993). Altruism and economics. *The American Economic Review*, 83(2): 156-161.
Zeitlyn, D., (2003). Gift economies in the development of open source software: anthropological reflections, *Research Policy*, 32: 1287—1291.
Wark, McK., (2004). *A Hacker Manifesto*, Harvard University Press, Cambridge (MA).
Williams, S., (2002). *Free as in Freedom*, O'Reilly, Sebastopol (CA).

THE EMERGENCE OF ALGORITHMIC MAN

THE NON-SELF-CORRECTIVE CONSCIOUSNESS

EVARISTUS O EKWUEME

Boston College

Department of Philosophy

Chestnut Hill, MA 02467 USA

Tel: 617 552 8194 evaristu@bc.edu

INTRODUCTION

This paper will explore the cognitional consciousness of a normal information system user with respect to insights and further questions arising from insight. What kind of consciousness is to be expected from the advancement of information systems or under the pretext that computer systems know it all? Ignorance of the computer software and programs are quite normal. However, the ordinary user of information systems is reduced to a simple follower of instructions (algorithmic person) and not necessarily an inquirer or a thinker. Routine jobs like entering information into a database in a supermarket does not require

any extraordinary inquiring mind. Let us take a simple example of a bank teller and customer:

Customer: "I would like to withdraw $200 from my account. Account #123456789"
Teller: (Enters the Account #): "I am sorry, Sir, the computer says you are bankrupt."
Customer: "That is not possible. I deposited $3,000 three weeks ago."
Teller: "I am sorry, Sir, the computer says you are bankrupt."
Customer: "Please check again! It must be a mistake"
Teller: "I am sorry, Sir, the computer says you are bankrupt."
Customer: "Did you understand what I just said?"
Teller: "I am sorry, Sir, the computer says you are bankrupt."

The teller operates on the simplest level of common sense, that of reading and following instructions without any regard to the counter evidence presented. The teller does not ask further questions that could reveal the actual status of the customer's account. He/She simply repeats the dedicated instruction for a "system error", "I am sorry, Sir, the computer says are bankrupt."

If it is true that insight only occurs in the intelligent, how is insight possible in a person who simply follows instructions? Thus, a high level of stupidity should be expected in the digital age because of the use of information systems that does everything for you, even "thinking"194. As Bernard Lonergan argues that the question of "pure unrestricted desire to understand" is human. However, the development of expert information systems that make decisions based on a certain knowledge-based programming raises questions. In working with such an assumption, the inquirer is already restricted in his human ability to seek insight through unrestricted questions. This inability to question one's reality is the denial of the human vocation as a subject[195]: "The question reveals your questioned experience as a possibly intelligible experience which your expected insight, when it occurs, will transform into an actual intelligible experience."[196] This intelligible experience will raise further questions to dynamically re-examine the experienced, the understood, the judged and the decided. The inhibition of this self-correcting consciousness experiencing, understanding, judging and deciding of one's own experience, understanding, judgment and decision are manifestations of the dramatic and individual biases.

---

194 This does not suggest that there are such machines already or that the Turing Test is the yardstick to determine such thinking abilities. However, this paper works on the assumption that the ordinary individual assumes such thinking capabilities on a machine whether it in fact is or is not capable of such an activity.

195 Paulo Freire, Education for Critical Consciousness, (New York: Continuum, 2000), 15.

196 Joseph Flanagan, Quest for Self Knowledge: An Essay in Lonergan's Philosophy (Toronto: University of Toronto Press, 1997), 17.

EMERGENT PROBABILITY AND SURVIVAL OF INTELLIGENT INFORMATION SYSTEMS

The digital age raises lots of questions on the biases. The main idea of asking questions is to seek answers. However, if about 99% of the answers are already available, the interest in questioning will be discarded. This state of human inquiry is better explained by Lonergan.

> According to Lonergan, biased courses of action that evade intelligent self-correction initiate downward spirals of decline, degradation and destruction not only of natural but also of cultural environments. Biases and decline have their own "logic" – the logic of a vicious cycles that lead to great destruction, unless something acts to reverse their downward trends (1992, 214-23, 242-63).197

The only possible intervention in the conditioned schemes of recurrence and the emergence of intelligent information systems are human intelligent self-corrective inquiry and action. In the digital age a deadlock of inquiry is possible should the information systems emerge as independent cycles. The destruction of humanity, at least as an inquirer or a knower, is possible. Unfortunately, given that the decline has already started, is an indication that it has already started setting conditions for "downward spirals of decline" within a certain interval of time. The conditions that have emerged are the wide use of expert and information systems, also the emergence of educational systems in grade, high school and universities that attempt to condition the intellectual impulse of students. The idea of using these systems is to make things physically less burdensome. The common phrase that accompany such education is, "All you have to do is...." What usually follows are instructions or prescriptions which the student have to memorize or simply follow "automatically".

Schemes of recurrence are fundamental to understanding the very notion of emergent probability. Lonergan argues that the schemes might be related by series of conditions such that, "if A occurs, B will occur; if B occurs, C will occur; if C occurs, … A will occur"198. There could be complexities in terms of the series or the conditions that need to be met before the hitherto following event will occur. The complexity might include a complete arrangement such that B cannot occur without A, and C without B. The probability for the occurrence of a whole set of

---

197 Patrick H. Byrne, "Ecology, Economy and Redemption as Dynamic: The Contributions of Jane Jacobs and Bernard Lonergan" on http://www.nd.edu/~ecoltheo/text_byrne.htm (Retrieved in August, 2 2004).

198 Bernard Lonergan, Collected Works of Bernard Lonergan: Insight. Ed. Frederick E. Crowe and Robert M. Doran. (Toronto: University of Toronto Press, 1992), 141.

events is (1/a x 1/b x 1/c x 1/n..) which is the product of the probabilities. However, if the schemes of recurrence is reflexive, it follows that the occurrence of one event will lead to the occurrence of other events and the probability is as follows (1/a + 1/b + 1/c + 1/n..). The idea of emergence rests on this profound shift in the probabilities. It could also be an elimination of the first series of events, for instance, "if F occurs, then G occurs; if G occurs, then H occurs; if H occurs, then F is eliminated"199. This is defensive circle. This circle explains the elimination of the certain events of schemes with low probabilities of survival, given the emergence of certain other schemes.

An understanding of the schemes of recurrence further explains that of the conditional series of schemes of recurrence. In conditional series of schemes of recurrence, the schemes, for instance P, Q, R for a conditional series "all prior members of the series must be functioning actually for any later member to become a concrete possibility"200. Thus, the series of scheme P will be functioning without the existence of Q and R, and Q can function without the existence of R. However, the series of schemes Q cannot function without P already in existence, and R cannot function without Q already in existence. Events have different probabilities of emergence and that of survival. The probability of emergence "consists in the sum of the respective probabilities of all the events included in the scheme and it arises as soon as the prior conditions for the functioning of the schemes are satisfied"201. There is also the probability of the survival of the schemes of recurrence. It is only assured if and only if there is a "non-occurrence of any of the events that would disrupt the scheme"202. Thus, the emergence of a scheme only highlights the probability of the emergence of later schemes, and also the conditions for the survival of the schemes already functioning. Emergent probability, therefore, is engendered by the summation of the conditioned series of schemes and their respective probabilities of emergence and survival.

The emergence of expert systems or "intelligent" systems as schemes in series of schemes of recurrence is setting the conditions for the emergent probability and possible survival of later schemes. It is setting the conditions for the emergence of human biases in the use of information systems. Expert systems or "intelligent" systems are not free of human biases either in their programming or in their use. If "intelligent" systems are conceived as independent schemes of recurrence in human environments like water cycles that are operative independent of human

---

199 Ibid.

200 Ibid, 142.

201 Ibid., 144.

202 Ibid.

conditioning, it follows that human intervention will be pointless. That is to say, if it is possible for super computers to produce other systems of their own choosing, install and execute these systems, independent of human direct intervention, then the questions of bias might not be raised. However, creativity and freedom of inquiry of the knower will become impossible, if the information systems become an independent scheme of recurrence, if they record, analyze and decide the course of action of the development of technology, economics, social and political aspects of human existence, then the human being will be only an event in the series of conditioned schemes. The conditions for the emergence of such independent information systems schemes of recurrence are already emerging in the digital culture. The idea is to develop information systems that follow the world process of emergent probability. However, the probabilities of emergence of such a complete cycle of recurrence and its survival will undoubtedly depend on human consciousness of such emergence and a deliberate intervention to shape its emergence and survival. It is possible that a development of such independent intelligent system could function like the rest of nature. Human imperfection and bias cannot be circumvented even in such complex systems.

The schemes of recurrence of information systems are non-existent without human programmers and users. Thus, the major question of bias in relation to information systems and commerce is the attribute of independent schemes of series of recurrence to information systems. If human beings surrenders their natural ability of seeking insight through experiencing attentively, understanding intelligently, judging reasonably and acting responsibly, but only follows computer instructions, it is a clear indication of a dramatic bias already operative in the use of information systems. Dramatic bias is a deliberate blind spot that blocks our questioning of our experiences and further questions that arise until we arrive at an insight. The supervening act of understanding which is the unrestricted and deliberate act of authentic inquiry into experiences, understanding and judging is the only way to over-come dramatic bias. As Kenneth Melchin put it, “The intelligibility grasped in an insight, the truth affirmed in a judgment, and the value realized in a decision to act all emerge in dynamic processes whose operations are identical in structure to probably emerging world process”203. For instance, it is common to hear in a business enterprise that “We can’t do anything about it, it is a computer problem.” In the example cited earlier, a teller informs a client, “I am sorry sir the computer says you are bankrupt.” In a situation whereby this is truly false, should the client go home? Such dependence on an information system could lead to further

203 Kenneth R. Melchin, History, Ethics and Emergent Probability: Ethics, Society and History in the Work of Bernard Lonergan. New York: University Press of America, 1989. 112.

problems if there is no self-correcting consciousness of the human person, the ability to ask further questions about the “intelligent” system and its potential inadequacies and incompleteness.

EMERGENCE OF THE ALGORITHMIC MAN (NON-SELF-CORRECTING COGNITION AND BIASES)

This dramatic bias is the beginning of a long cycle of decline of the human intuition and creativity. Lonergan argues that “Deep within us all, emergent when the noise of other appetites is stilled, there is a drive to know, to understand, to see why, to discover the reason, to find the cause, to explain”204. It is possible in the digital age to blindly follow the dictates of a computer instruction or error message, than to inquire further into the possible explanations: “Lonergan’s account of emergent probability in the human order incorporates the fact of human failure to consider questions raised by their endeavors, failures to seek answers even to all the questions they do raise, and refusals to act according to what they come to understand as the best courses of action.” It will not be an exaggeration to say that swiftness of information access in businesses and institutions could lead to a decline in human inquiring. And thus, the emergence of the algorithmic person. The computer systems become the god of the digital age, whose dictates requires simple user inputs of “yes” or “no”. The systems become sole authorities that are too complex to understand and too abstruse to contradict. The individual might be intelligent enough to understand the common institutional schemes.[205] However, the individual is not intelligent enough to recognize and acknowledge himself as being diminished by his actions. The individual bias is accomplished by suppressing fellow feelings and spontaneously triggering falsely interconnected insights according to the prescriptions or algorithms already internalized as absolute solutions to any given problem.

To illustrate the non-self-correcting process of the algorithmic man and the natural dynamic self-correcting cognitional structure of the human mind, let us take an example of computer troubleshooting. Working as a network administrator in a small academic institution, I realized how frustrated students became when they have simple computer malfunctions. You probably have experienced this yourself sometimes. I would advice the students to go to the “Help Menu” of any program to fathom the solution. The troubleshooting programs are designed to do the questioning and the user simply follows the

---

204 Lonergan, 28.

205 Flanagan 81

instructions and chooses any options. Given that the individual does not understand the general computer system and particularly the problem at hand, he will be frustrated. The format for these “Help” instructions are usually as follows: “What problem are you having?” What follows are series of choices for the user, and a “Next” button to advance to the next level of troubleshooting. This general question is not exhaustive of the possible questions that a student might have. The minimum level of troubleshooting instructional questions that most ordinary users can stand are four. At the fifth or sixth level of questioning they become frustrated. The major cause of this frustration is the fact that the individual is not the one asking the questions. He is given the questions and “all he has to do” is choose a cause of action. This form of troubleshooting a computer problem does not take into cognizance the self-correcting cognitional structure of the individual. The individual naturally wants to ask questions and also to make the choices and if possible increase the possible choices in order to make the best decisions. In one instance, a student said to me, “Why don't I just ask my questions and the computer can give me answers?” The student intended to engage his consciousness and the self-correcting activity of his mind. The student wants to experience the tension of inquiry that is constitute of insight and not amputate his natural ability to understand the problem. He intends to heighten his consciousness. According to Lonergan, the only way to objectify one's consciousness is by heightening it. It is something each individual has to intentionally and consciously do for himself and cannot be achieved by following already constructed questions and answers:

> To apply the operations as intentional to the operations as conscious is a fourfold matter (1) experiencing one's experiencing, understanding, judging and deciding, (2) understanding the unity and relations of one's experienced experiencing, understanding, judging, deciding, (3) affirming the reality of one's experienced and understood experiencing, understanding, judging, deciding and (4) deciding to operate in accord with the norms immanent in the spontaneous relatedness of one's experienced, understood, affirmed experiencing, understanding, judging, and deciding.[206]

The absence of the effort to experience, to understand, to judge and to decide is constitutive of stupidity. This is the case with the algorithmic man who simply follows the troubleshooting instructions without prior questions that are seeking answers. The algorithmic man manifests a static or non-self-correcting cognition that neither seeks to experience, to understand, to judge and to decide but is solely dependent on the troubleshooting instructions. The four different levels of consciousness that defines and differentiates individual from natural or artificial

206 Lonergan, Method in Theology, 14-15.

inanimate and animate objects are not possible in the non-self-correcting consciousness. For Lonergan, there are four levels of consciousness; the empirical level which consists sensing, perceiving, imagining, feeling, speaking, moving; intellectual level which consists of inquiry, understanding, expression of the understood, presuppositions and implications of what is expressed; rational level which consists of reflection, marshaling out evidence, passing judgment on the truth or falsity, certainty or probability of a statement; responsibility level which consists of concern for self, one's own operations, goals, and possible courses of action, evaluation of the course of actions, deciding and carrying out the decisions.[207] The non-self-corrective consciousness does not advance beyond the first level of consciousness. This level of consciousness is submerged and lacks the "ability to intervene in reality as it is unveiled."[208] The algorithmic man lacks "*conscientização,* the deepening of the attitude of awareness characteristic of all emergence."[209] It is only through inquiry that the individual can be truly human and true knowledge only emerges out of restless and impatient intervention and re-intervention in and with human experiences.[210]

ALGORITHMIC MAN OF THE DIGITAL AGE

> In order that such a man also be ruled by something similar to what rules the best man, don't we say that he must be the slave of that best man who has the divine rule in himself? It's not that we suppose the slave must be ruled to his own detriment, as Thrasymachus supposed about the ruled; but that it's better for all to be ruled by what is divine and prudent, especially when one has it as his own within himself; but, if not, set over one from outside, so that insofar as possible all will be alike and friends, piloted by the same thing[211] (590b).

In this scenario, the spirited man, the guardians and the merchants will be better ruled by the philosophers. This centralized intelligence in the government of everything else is the core of Plato's idea of the philosopher kings. Every member of the *polis* should avail themselves the choice of knowing what is best for himself. This is the job of the philosopher kings who will issue the appropriate instructions to

---

207 Ibid, 9.

208 Paulo Freire, Pedagogy of the Oppressed , Trans. Myra Bergman Ramos. (New York: Continuum, 2000), 90.

209 Ibid.

210 Ibid.

211 Plato. The Republic of Plato, Trans. Allan Bloom. 2nd Edition, (Chicago: Basic Books, 1991), 273.

everyone's advantage. In the digital age the centralized intelligence are the computer systems and their programmers and everyone's else avails himself of the potential use of his intellectual capabilities. In such a case, the deposit of cognitional self-correcting consciousness is of no practical or moral use to the individual. The prescriptions of the information systems and their programmers are sufficient for human existence. Such modern science, as Lonergan points out, wants to be a natural science that can predict every human action while at the same time exhibiting the free choice of individuals. This is an estrangement of man's actual reality from man: "Man sets up an inhuman order because he conceives man as a component in a machine; and man hates that machine."[212] The non-self-correcting component of an information system that is devoid of meaning and insight. This level of consciousness and commonsense can easily be replaced by another component without any conceivable practical difference.

As Paulo Freire points out this is a "'domestication' of man's critical faculties by a situation in which he is massified and has only the illusion of choices."[213] The individual is "Excluded from the sphere of decisions being made by fewer and fewer people, man is maneuvered by the mass media to the point where he believes nothing he has not heard on the radio, seen on television, or read in the newspaper. He comes to accept mythical explanations of his reality. Like a man who has lost his address, he is 'uprooted'."[214] Assistencialism, therefore, is the inability of the individual to participate in a historical process as a subject. It enforces the denial of one's experience, understanding, judging and deciding, while offering inattentiveness, unintelligibility, irresponsibility and unreasonableness, and most especially passivity and anti-dialogue.

GROUP BIAS, GENERAL BIAS AND THE LONG CYCLE OF DECLINE

Individual biases do manifest in groups. In the digital culture, the programmers and their sponsors form an elite group in collaboration with the educational systems. In group bias, group loyalties are fostered to repress relevant questions that could engender understanding.[215] They are offered elaborate excuses and rationalizations that people do want things simple and easy. They do not need to think for themselves or rather prefer someone do the thinking for them. This group egoism

212 Bernard Lonergan, Method in Theology. (London: Darton, Longman and Todd, 1972), 46.

213 Paulo Freire, Education for Critical Consciousness, 34.

214 Ibid.

215 Flanagan, 84

generates a socially supported ideology that information systems "know everything." There also the formation of a group of ordinary users or community of users. These groups are habituated into inattentiveness, unintelligibility, irresponsibility and unreasonableness through education. This kind of education is what Paulo Freire refers to as "banking education."[216] The task of the education is to "'fill' the students with the contents of his narration which are detached from reality."[217] What is "loaded" into students are basically prescriptions and algorithms of what to do and say. Their naïve consciousness is only meant to execute those algorithms as they are loaded without modification. The students patiently receive, memorize and repeat the algorithms. This form of education is founded on the general bias that the recipients are mere objects, adaptable and manageable beings. They are as good as other programs loaded into the memory of a computer system. They can only do what is prescribed. The group bias of banking education minimizes or inhibits the students creative power to ask questions and to be fully human.[218] Freire quotes, Eric Fromm's *Escape from Freedom,*to best illustrate the cognitional state of the modern man: "[Man] has become free from the external bonds that would prevent him from doing and thinking as he sees fit. *He would be free to act according to his own will, if he knew what he wanted, thought, and felt. But he does not know. He conforms to anonymous authorities and adopts to self which is not his.* The more he does this, the more powerless he feels, the more is he forced to conform."[219] This pertains to his ordinary life and his industrial job. Lonergan quotes Karl Jasper to also depict the cognitional state of the modern man in an era of technological possibilities. The modern man is cognitively passive, "everything connected with his job had been worked out for him by somebody else; he just goes through the motions. And the fellow that works it out is in the same position."[220] There is not room for personal creativity or critical consciousness, thus, there is no room for persona decision or achievement. In a complete state of ignorance or partial abandonment of inquiry, the individual simply follows instructions blindly given that he does not grasp the situation due to inattentiveness, unintelligibility, irresponsibility and unreasonableness.

Personal and group biases originate from a failure in the common-sense knowing. However, the general bias is a failure "on the part of practical knowers to accept the fact that common-sense knowing is a limited specialized form of knowing."[221] The problem of general bias stems from

---

216 Freire, Pedagogy of the Oppressed, 54.

217 Ibid. 53.

218 Ibid, 57.

219 Freire, Education for Consciousness, 6. Italics are added for emphasis.

220 Lonergan, Method in Theology, 45.

221 Flanagan, 85.

the tension between interested and disinterested knowing. The commonsense knower is not historically conscious and is preoccupied with short-term problems without historical bases. What generally ensures from general bias is the long cycle of decline. The individual and group biases result into short-term disorders, however, general bias are long term disorders that stem from what was completely neglected by different groups, in this case, the programmers and the ordinary users. As pointed out above, the idea of designing computer information systems is to ease the burden of intellectual activities of the users, given that the choices are preprogrammed. The long cycle of decline is generated from group and general biases which lead to series of lower viewpoints. The dialectic is "between interested practical knowing of the commonsense and the disinterested desire of theoretical knowing."[222] This dialectic inhibits progress. A possible resolution of this situation is that of cooperation between the two groups. To overcome a long cycle of decline, a new and higher viewpoint that can deal with the problem of non-self-correcting cognition at its source is needed. A critical and self-correcting higher viewpoint that is historically conscious of the root of the problem of non-self-correcting cognition. This way the future information systems, educational systems and practical existences of the new generations will be designed not with empirical, intellectual, rational and responsible levels of consciousness that promotes full human consciousness.

CONCLUSION

> Perhaps the greatest tragedy of modern man is his domination by the force of the myth *that computers and their programmers know all).*
>
> Gradually, without even realizing the loss, the ordinary user relinquishes his capacity for self-corrective experiencing, understanding, judging, deciding; he is expelled from the orbit of decisions.
>
> Ordinary user do not perceive the tasks of the time; the latter are interpreted by an "elite" *[computer programmers] and* presented in the form of recipes, prescriptions or algorithms.
>
> And when men try to save themselves by following the prescriptions, they drown in leveling anonymity, without hope and without faith, domesticated and adjusted.

---

222 Ibid, 86.

This domesticated consciousness is the algorithmic man[223]

Insight is not possible through following of algorithmic computer instructions. The emergence of computer technology sets new conditions for the probabilities of emergence and survival of human cognitional structures. Thus, the development of intelligent information technology and systems that set up a complete cycle of schemes of recurrences present difficult challenges for the human knower. The possible biases that accompany the design and development of information systems, coupled with the consequential biases of non-self-corrective existence have adverse effect in human understanding. Despite the biases, however, information systems could be a possible solution to documenting, analyzing and understanding the different patterns and schemes of recurrence operative in technological, economic, social, political and religious dimensions of human history. In as much as the conditions that need to be satisfied for the probabilities of emergence of a certain event or scheme can be understood, such insight or information could prove valuable to the improvement of humanity in general.

Definition of Terms

Algorithmic Person: It is an ordinary user of a computer system who simply follows instructions as dictated by the computer without further inquiry in the possible incompleteness of the system. This person inhibits the self-correcting mechanism of the human dynamic cognitional structure.

Insight: It is the supervening act of human self-corrective understanding

Emergent Probability: It is the probability of emergence of series of events from different schemes, such that the emergence of one event sets the probable condition for the emergence of other events or series of events of another scheme.

Probability of Survival: It is the tendency for an emergent event, scheme or series of schemes to endure after emergence as long as the conditions that will warrant its elimination is not reached in time.

223 Freire, Education for Critical Consciousness, 6. This is adopted and paraphrased from Paulo Freire's conception of man as object that is submerged in history. Man in this situation attempts to rescue himself by following the instructions of those who oppress but only becomes domesticated and adjusted to the misguided world he fights against.

Schemes of Recurrence: It is the tendency for schemes to reoccur as long as the conditions continue to be favorable.

Conditional Schemes of Recurrence: In conditional series of schemes of recurrence, the schemes, for instance P, Q, R for a conditional series "all prior members of the series must be functioning actually for any later member to become a concrete possibility."

Scotosis: This is individual dramatic bias in which the individual deliberately interferes with the natural cognitional structure of questions and further questions leading to insight. There is individual, group and general bias that are common sense biases that fail to follow the cognitional path towards insight.

References

Bernard Lonergan. Collected Works of Bernard Lonergan: Insight. Ed. Frederick E. Crowe and Robert M. Doran. Toronto: University of Toronto Press, 1992.

Bernard Lonergan, Method in Theology. (London: Darton, Longman and Todd, 1972)

Gordon Steven R and Judith R. Gordon. Information System: A Management Approach. New York; Harcourt Brace & Company, 1996.

Joseph Flanagan, Quest for Self Knowledge: An Essay in Lonergan's Philosophy.
Toronto: University of Toronto Press, 1997.

Kenneth R. Melchin, History, Ethics and Emergent Probability: Ethics, Society and History in the Work of Bernard Lonergan. New York: University Press of America, 1989.

Patrick H. Byrne, "Ecology, Economy and Redemption as Dynamic: The Contributions of Jane Jacobs and Bernard Lonergan" http://www.nd.edu/~ecoltheo/text_byrne.htm (Retrieved in August, 2 2004).

Zwass Vladimir. Foundations of Information Systems. Boston, McGraw-Hill, 1998.

Bocij Paul, Dave Chaffey, Andrew Greasley, Simon Hickie. Business Information Systems: Technology Development and Management. Ed. Dave Chaffey. London, Pitman Publishing, 1999.

Plato. The Republic of Plato, Trans. Allan Bloom. 2nd Edition, (Chicago: Basic Books, 1991), 273.

Paulo Freire. Pedagogy of the Oppressed , Trans. Myra Bergman Ramos. New York: Continuum, 2000.

Paulo Freire. Education for Critical Consciousness, New York: Continuum, 2000.

RENAISSANCE SEARCH ENGINE AS A COGNITIVE ARTIFACT

Christopher YUKNA (yukna@emse.fr, yuknachris@yahoo.com), Ecole des Mines de Saint-Etienne, FRANCE
Fabrice MUHLENBACH (fabrice.muhlenbach@univ-st-etienne.fr), Laboratoire Hubert Curien (ex-EURISE), Université Jean Monnet, FRANCE

Introduction

The Gutenberg printing press was the motor for the Renaissance. The scholars of that time could know almost everything about many topics. In a very similar way the Internet and omnipresence of computers is changing our society. What will come from the enormous development of informational technologies is still to be determined. However, with search engines, web alerts, newsletters, forums, and many other indexes and databases on line, society seems to be laying the groundwork for a CyberRenaissance. Would it be possible to develop consciously a cognitive artifact or artifacts to allow the reemergence of the Renaissance Man? We propose a solution that permits scientific advances by neogeneralists in the footsteps of Erasmus, Francis Bacon and Leonardo Da Vinci. Intrinsically, the Internet may be similar to a printing press on speed, but a computer is not a book. For the CyberRenaissance, these differences will be significant.

The Historical Renaissance Premise

Conditions similar to those in the Renaissance exist today. History repeats. If you examine the cultural, philosophical, political aspects of the Renaissance and what followed, it may help you to understand the even

more rapid and pervasive changes which are before us in the 21st century.

Characteristics of the Historical Renaissance

The Renaissance is the appellation of the historical period usually considered to have begun in the 14th century in Italy and the 16th century in the northern Europe where impressive changes in artistic and scientific activities have fostered the transition between the end of the Middle Ages and the industrial revolution. The re-birth –etymological meaning of the French word “Renaissance”– was associated with the rediscovery of the ancient Greco-roman heritage and the absorption of scientific knowledge from Arabic civilization. The major factor of this rediscovery was the circulation of information, starting with the collaboration between artisans and artists in the Italian cities, the assimilation of foreign influences coming from the trading ships, the introduction of the changes in France after the invasion of Italy, and the diffusion of printed material to the whole Europe by the Gutenberg press (Hale, 1993).

The Importance of Libraries

Renaissance libraries were crucial to the scientific and cultural philosophy of the day, in particular the Papal Library. Its Pope, Clement V, was an obsessive collector of books. While perhaps not true, under his patronage the Vatican Library seemed to have most of the books ever published on its shelves. Image entering his library and seeing a wall of books three or four meters wide. This is where all the books of science or close to it would be found. You could in your lifetime have read them all. Although, you might have had to learn Greek or several other languages to do so. Certainly, if you needed to know something in say Biology you could have browsed though all of the indexes on that shelf and thus find the chapter you needed. Later in history card catalogues would have helped you do so, but at a much grander scale.

Today, things have changed significantly. How long would be an aisle dedicated to all of the books and magazines of just one discipline: Chemistry? No librarian has that kind of budget or space. Incredibly, it is unlikely that you could read all of the Science that is published worldwide in a single day, even if you spent your entire life trying to do so. Is it any wonder that with so much information to deal with, young researchers are rushed into specializing so that they can get a handle on what is happening in their field before they are too old. Yet in the Renaissance, everyone was supposed to dabble in the Arts and Sciences and arguably society was better for it. Where are our cyber analogies to the libraries of the Renaissance? Can we have all of Science at our fingertips?

The Change of Point of Views and its Spread Worldwide

There were many inventions and creations in arts and sciences during the Renaissance. The rediscovery of experiments and mathematical studies elaborated during the Roman period or the Middle Ages, added to philosophical aspects of the human being and his link to the world, have been essential to develop the perspective theory (Panofsky, 1965). The place of individuals in the space changed the traditional way of representing people in paintings and drawings. Before the Renaissance the size of objects and characters were not drawn according to distance away but to their symbolic importance.

The first theorization on perspective were made by Arabic mathematicians working on optics around the year 1000 AD but it is only centuries later that Italian artists began using algebraic methods of perspective in their drawings. From Florence, the perspective paintings and the associated techniques were proliferated in a few decades throughout Western art.

The German artist Albrecht Dürer, after his visits in Italy, that he became a major force in the distribution of perspective via his tutorials. Dürer, a archetypical Renaissance man, was not only a painter, wood carver and engraver, but also a mathematician. Combining knowledge in arts and science, he wrote a tutorial where he explained his theoretical work on geometry and perspective, proportion and fortification, just like today where web tutorials on any domain can be found on the Internet.

The Renaissance Men

The Renaissance saw the appearance of men who excelled in many fields. A Renaissance Man, or "polymath", is characterized by being someone who bridges boundaries of disciplines, disproves existing paradigms, and takes advantage of techniques and concepts in other fields to use in another. During the Renaissance, the complete savoir-faire of science and art existed in a limited number of books. In such a way, it was possible to have a holistic view of the state-of-the-art of any subject.

This holistic view of this knowledge is still a valid approach. Scientific discoveries, and moreover scientific revolutions, happens when the disciplinary gates are broken. Studying a subject from only one perspective often fails. Take the parable of touching blindly an elephant: the animal could be considered as a tree (leg), wall (side), snake (trunk), rope (tail) or even stone (ivory tusks) if only one part of the animal is felt.

A new Renaissance?

Internet and the Gutenberg Press

The Internet is having a similar deep impact on work, leisure, knowledge and worldviews as the printing press of the Renaissance. Current scientists are using the Internet to delve for information in their research areas, famous scientific publishers offer an electronic format of their reviews on the Web, book publishers via Amazon and GooglePrint allow access as well to their materiel, and thus the stock of information on Internet is becoming ever more available, rapid and accessible, just like the huge change during the Renaissance which was the explosion of the printed word as compared to the hand-written word in monasteries.
How are the printed page and the Internet divergent? A lot has been written about hyperlinks and hypertexts and how they are fundamentally different than the printed page. The knowledge acquired by the two methods is also dissimilar.
Traditional and computer-based learning are also divergent. Learning from a book or a predefined and structured course lends itself to linear thinking and specialization. However, entering cyberspace and goal oriented learning requires a much more generalist approach.
We would like to make a visual analogy. Let's compare learning to downloading or uploading. Information acquired by reading a book is like a streaming video or downloading software. Anyone familiar with the environment of the web knows what happens when you download: a horizontal bar appears on the screen. This horizontal bar is slowly filled from left to right. When it is finished the space is filled and the download complete. This is how we picture information transfer with a book. For the Internet/computer experience it seems more like downloading with a peer to peer system like Kazaa or Emule. There is still a horizontal strip, but the space in the bar is not filled from left to right. Small vertical lines appear throughout the space and little by little the bar is filled in. The P2P experience is generally more adapted to extremely large downloads or if the analogy holds extremely broad types of learning.

The Renaissance Men of Today

Notice that many innovative recent works are only transfers and adaptations of the concepts or procedures produced previously in other domains, not from sciences to arts like in Renaissance, but more from hard sciences to humanities, or inside hard sciences.
Here are just a few examples. Herbert Simon (1916 – 2001) is well-known to be a precursor of artificial intelligence. He was awarded the

ACM's Turing Award along with Allen Newell in 1975 for his contributions to artificial intelligence and the psychology of human cognition. Three years later, he was awarded the Nobel Memorial Prize in Economics! More recently, in 2002, the psychologist Daniel Kahneman, famous for his work on the cognitive basis for common human errors with Amos Tversky, who won the Nobel Prize in the same area for his work on behavioral finance and hedonic psychology. Obviously, mixing disciplines is fruitful, like in cognitive science where to study scientifically the mind, uses the fertile collaboration between psychologists, computer scientists, linguists, philosophers and neuroscientists.

Dealing with the CyberRenaissance

Access to Information: Virtual Libraries

Through the eyes of Google *et al*. research is quite different from just a few years ago. Return to that imaginary aisle of Chemistry books mentioned before. If this total Chemistry section existed and you went looking for some information chances are that the book, paragraph or pages that you need would be buried somewhere down a long corridor overflowing with tomes or journals. The time you would spend searching would be prohibitive.

Even though at present search engines are limited to rather old methods of searching: Indexing by copying. Text based search works surprisingly well if you can generate the right keywords. With the right incantation, Google or one of its rivals will take you to not just the book or chapter but the right paragraph or sentence. Which is a great deal better than wandering around staring at kilometer long shelves of books.

There are limits. In very real sense everything on the web is encrypted and we are constantly hacking with Google to find what we want. This is not to deny the power of phrase searching. Amid the six billion pages indexed in Google it is possible to find almost any page with just two or three words. Of course search engines will continue to evolve and so will the Internet. Search is being accelerated.

Most researchers would rightly scoff at the idea of the Internet as being a reliable source of scientific information. Certainly, it is a nice place to go to look for recipes or gardening tips but Science no. The Internet has no order and is even rife with pseudoscience. Yet, one surprising source of credibility is the collectively reviewed phenomena like Wikipedia; whose articles have come to rival traditional Encyclopedias. Wikipedia and its related open source books and wikibooks are seemingly getting better and more sophisticated all the time. But maybe you can not wait.

Another solution to the integrity question is to use the specialized Science search engines like Google Scholar, Scirus, OAIster, etc., whose algorithms try to filter out much of the unverified. Unfortunately, a great deal of the responses are sites where you have to pay to view the articles, thus diminishing greatly the utility of this strategy. Many schools counter by subscribing to paid services which have most of the articles available (but not all).
This brings us to the Pirate Library. Young researchers in laboratories worldwide often need special articles or reference books to complete their thesis. Even well endowed schools do not always have the time or money to cater to the whims of grad students. So what are these young scientists doing? Conjure up that virtual chemistry aisle again. This time it exists on a student's hard drive in the form of thousands and thousands of searchable PDF files. How is it accomplished? The "conspirators" are not even organized, but resemble emergent behavior. People around the world take or create PDF books and articles then make them available via P2P networks, Emule Kaaza, etc. Quite often it is even simpler than that. Many companies like Rapidshare will let you upload a file for free and generate a link where others can down this content. To see if someone has stocked the book you want: Simply search in the web for the title or author plus the name webspace provider and quite often you will find an entry in a forum or newsgroup which points to the book or science article available for download. The average 700 page reference book takes up about 15 megabytes in PDF format. It has been eye opening experience when visiting labs where researchers have stocked one to two hundred gigabytes of science books and articles on one machine, all accessible to some form of desktop search. This technique vastly levels the playing field. Third world institutions can amass a better database than the richest university. Those who use the Pirate Library justify their treasure trove by saying that they will never actually read this texts but merely browse or consult them. In addition, they claim that any research papers they write will be open source for the world to use. Like after the Renaissance, revolutions will happen.

Data, Information, Knowledge… Wisdom?

Different problem from Renaissance time is that today we are overwhelmed with information. Child of statistics, information retrieval, machine learning and pattern recognition, data mining and knowledge-discovery in databases are the tools and solutions to address this problem. By definition, data mining is "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley, Piatetsky-Shapiro, and Matheus, 1992), something you can not apply to the printed word.

The goal of data mining is to transform raw data into meaningful information, and information into useful knowledge, and even wisdom. Because the most important source of easily accessible information is now on the Internet, it is essential to possess data mining tools to explore the web and extract not only answering simple questions but pointing out a network information by cross-referencing web pages. Could we develop a sort of catalyst to enable the Renaissance man to flourish today? Using data mining principles which try to increase the level of information, from raw data to information by gathering and connecting of parts, from information to knowledge by formation of a whole, and even maybe from knowledge to wisdom by joining the wholes, could we facilitate a new humanism?

Reasons for Generalists: Chinese Room reloaded

In the Renaissance, we have explained that there was a holistic scientific approach. Could this return to the mainstream? There are some mega trends or indications that this may come to pass. We live in a civilization where the work of specialized professional teams is increasingly done at home by amateurs sometimes with specialized devices or well configured computers. An entire recording studio is available on a PC or a Mac. Where once you needed a complete laboratory complete with PhDs and assistants to test for pregnancy, now women do that by themselves in the bathroom. Hollywood special effects were done in a kitchen with a group of actors in Finland (Star Wreck: In the Pirkinning, http://www.starwreck.com)

Here lies the critical question: Can it be possible to copy and paste a series of arguments and works from divergent fields and coherently assemble them into a breakthrough theory? Is it really necessary to understand all of ecology, high energy plasma, and medicine to work in an interdisciplinary fashion with these domains? It seems within the realms of possibility that even today, without any sophisticated cognitive aids, by skimming the surface of a subject enough insights can be gleaned to produce interesting and viable scientific work. Seeing the latest results without any preconceived ideas should lead to new perspectives and theories. Now let's postulate potent cognitive aids to facilitate this task. Which each new version and as the operators better configure the program to the endeavor, the capacities of this human/computer cyborg should increase.

Which brings us to a slightly more controversial concept: Can generalists do significant work in specialized fields? Perhaps an analogy is in order. Let us paraphrase John Searle's Chinese room experiment. Consider a subject that you have never studied, for example astrophysics. To you, at least initially, astrophysics is meaningless. Now instead of a rule book on Chinese squiggles, you have a computer interface. This is capable of

rendering comprehensible or at least manipulable astrophysics. With this very astute interface you are able to formulate and test theories perhaps like some ultra simulator, setting up conditions on stellar formation and then analyzing and evaluating the results. Suppose that the interface permits you to do work that approaches that of an astronomer. At what point are you a real scientist, if ever? Maybe your work in this field will surpass the professionals in a way. After all, companies pay consultants for their outside viewpoints. These generalists will make mistakes. But there is a necessity for making errors. With perfectly replicable DNA Evolution could not exist.

Man and Machine Interface

As mentioned earlier most search is text based. Moreover, for the majority of the researchers around the world these magic words are in English and not in their mother tongue hence in some small way handicapping them. Assume that you see an object or photo at a conference and do not know its name. How to recover an image from the web? Regardless, are search engines and their like moving toward a more organic relationship with their users? This seems to be the case. Nonverbal, search engines exist. While in, at best, beta version, these innovative tools point to a very different future for search. Take for example Retrievr. With Retrievr (http://labs.systemone.at/retrievr/) you try to draw the image that you are searching for. The results are amusing. Music search is beginning to incorporate notes (classical music search http://iwamura.home.znet.com/kbdif/kbdif.html) or tempo (SongTapper www.songtapper.com/). The possibility that your computer will use its webcam to assess your emotional response to search outcomes to further configure these same results is no longer in the so distant future?

The Renaissance Search Engine

The Need of a different Search Engine to help Researchers

We propose to develop the "Renaissance Search Engine", a cognitive artifact whose primary function to help future Galileos navigate and formulate their ground breaking ideas. Not everyone needs to become a polymath to function. However, science has excluded those who have not specialized and ironically this appears to be just the people needed for many actual research projects. We define this virtual artifact by what it needs. This falls back to what are the needs of its projected users. We identify three properties to such engine.

First Property Specific to the Renaissance Search Engine: Identifying Similarities

How could someone notice potential links between fields? Data mining and pattern recognition are answers. Through the history of science noticing similarities between examples in nature and scientific or technical endeavors has lead to numerous breakthroughs. With the overly specialized fields in science today, it is easy to see how two fairly unrelated domains could be studying the same phenomena under different names. Obviously, having a tool that could scan for similar phrasing in diverse fields would be a boon to research. Even experts can be blinded by their own semantics. Take just one example, a charged particle, cosmic ray, proton, ionized hydrogen atom, plasma, etc., could all be synonyms, depending on the situation. Science can get complicated. Even simpler solutions, small cognitive aids to help accomplish this manually are needed.

Second Property of the Renaissance Search Engine: Identifying the Pattern of Errors

Often, to be truly creative in science you must decide, at some point in time, that an existing theory or school of thought is wrong. More than the Occam's razor principle, which suggests when faced with two solutions to take the simplest, we recommend that sometimes, when an existing theory appears so complicated as to be an affront to reality, you should apply Occam's guillotine and reject entirely this theory. This is not as sacrilegious as it sounds. Many studies have shown that experts make as many errors as other people. There is also the very human trait not to question really important or long held beliefs. Often at business meetings, where to place the coffee machine can take hours while the expensive items on the budget pass without debate. In this case the Renaissance Search Engine (RSE) might be used to stimulate the imagination (Daniel Dennett, 1982) or doubt of budding Leonard Da Vincis. Therefore, providing some indicators of errors to budding iconoclasts would be appropriate.

Experts, when they make a mistake, start to make more and more involved explanations as to why their diagnosis is correct. Notice the case of Nicolaus Copernicus: The experts of the time, like Tycho Brahe, explained the geocentric universe by adding more levels of epicycles (circles within circles). Going from a simple model to a complex model without resolving the underlying problem might indicate that the theory is incorrect or at least here is a place to look. By developing methods to analysis the enormous scientific and cultural online content it is conceivable that potential errors might be identified.

Third Property of the Renaissance Search Engine: Spotting the Void

One intriguing use of the Renaissance Search Engine would be to locate empty research space. Obviously, the easiest way to do original research is to go where no one has gone before. Search engines are already being used to find nearly empty space of a sort... Googlewhacking, uses Google to point out where two words are not commonly found together on a page. The object of a googlewhack is to type only two words (which can be found in a dictionary) in Google and have only one page as the result.
Furthermore, when people go to Wikipedia and can not find an article they want, this often becomes the subject of their next entry in Wiki. Finally, before pursuing a new line of research, most scientists search the web for similar ideas so as not to redo existing work. However, identifying an abyss or gaps in knowledge is not as easy as it sounds. Let's take for example missing images. Let's postulate that while there are loads of photos taken by the Hubble telescope no one has made available on the web a photo of the space telescope itself. This would be fairly easy to spot at random by a human, but not at all clear how the Renaissance Search Engine could highlight this fact to its user. New forms of Science: Chaos or Games Theory, Evolution etc. are being developed all the time. Often they are applicable to most of science in general. Noticing where this cross fertilization has not taken place is also a necessary function of the RSE. Evidently more complex approaches could be developed to point out areas of little studied topics.

The CyberNautilus, a Renaissance Search Engine Prototype

It seemed a shame to write about the possibility of the RSE without creating at minimum a mockup. It would have been nice to have a huge budget for teams of programmers, cognitive scientists, search engine experts, etc. Creating some sort of expert system and neural net hybrid with AI agents and automatic user alerts would have been a snap. Alas, this was not the case. So we set our sights lower and used a few lines of javascript to launch several science related search engines at once. What emerged is the CyberNautilus. (http://yukna.free.fr/science/search/cybernautilus/cybernautilus.php)
Unexpectedly, this primitive decision support system works well. Opening up several different windows at a time and each with its own brand or style of results gives the breathtaking impression of sophisticated and incredibly quick searches. It is an illusion of intelligence, nevertheless. One feature that particularly helps the comprehension of science topics and research was the addition of yahoo images. Searching scientific terms in images often clarified the subject or led to more pertinent websites than via text. While being nothing more than a glorified mockup,

the CyberNautilus is a proof of concept. Using this webpage both sped up and broadened scientific searches.

Conclusion and Perspectives

In this paper we studied a simple premise: “We are in a historical era that resembles the Renaissance” and followed it to its conclusion. First, we attempted to report the similarities between today and the Renaissance. Second, the mega trends that appear to be transforming our society into a sort of CyberRenaissance were chronicled. In particular, the obvious differences in the philosophy as it applies to the Arts and Science was compared. Many parallels abound: the Internet can be seen as the Gutenberg Press, or the Vatican Library and search engines. A way to facilitate the holistic approach of yesteryear to today by considering the development of the Renaissance Search Engine. What type of characteristics and capacities would be needed to let an intelligent musician or artist suddenly tinker with science? The Pirate Library, a new and relatively secret technique among researchers was presented.

A simple prototype, the CyberNautilus, was developed to test these ideas. While extremely primitive this mockup seemed to be promising. Future work will include developing a much more elaborate Renaissance Search Engine, and applying some of the principles studied to concerns in Science and Medicine. The prospect of searching though all of the publications of Science at once is technically feasible. What remains to be done is its implementation.

Bibliography

Dennett, Daniel (1982), “Notes on Prosthetic Imagination”, *Boston Review* 7(3), pp. 3-7.

Hale, John (1993), *The civilization of Europe in the Renaissance*, Atheneum Publisher.

Frawley W., Piatetsky-Shapiro, G., Matheus, C. (1992), “Knowledge Discovery in Databases: An Overview”. *AI Magazine*, Fall 1992, pp. 213-228.

Panofsky, Erwin (1965), *Renaissance and Renascences in Western Art*. Stockholm: Almqvist & Wiksell.

Searle, John R. (1990), “Is the Brain’s Mind a Computer Program?”, *Scientific American*, January 1990.

Computing and Publishing: an Epistemological Issue

Roberto FELTRERO

Department of Science, Technology and Society

Institute of Philosphy, Spanish Research Council

0034 91 4117005 rfeltrero@gmail.com

Eduardo BUSTOS

Department of Logic,History and Philosophy of Science

Faculty of Philosophy, Spanish Open University (UNED)

0034 91 3986931 ebustos@fsof.uned.es

The widespread use of computational tools for scientific writing and publishing implies an important increase in the variety of possibilities for the collective production of information and scientific knowledge. Therefore, the epistemic consequences of this use should be acknowledged an important place within the realm of assessment parameters for technology and the new methodologies implemented through them. New ways of scientific publishing mediated by computers and Internet, have become a topic of ethical, practical and economical concern that, we propose, should be primarily regarded as epistemological.

We deploy here the notion of *epistemic site* as a tool for assessing the epistemological consequences of the use of computers and Information Technologies on activities of production, communication and distribution of information and knowledge. This concept includes a variety of evaluative parameters and principles that are useful to articulate epistemic and moral concerns about computational resources for publishing. The Open Access E-journals movement will be the chosen case study to be evaluated within these principles.

Keywords: Epistemic Site, Social Epistemology, Information Technologies, Internet, Open Access, E-prints, Self-Archiving, Scholarly Publishing, Intellectual Property

Introduction

The widespread use of computational tools for scientific writing and publishing means an important increase of the possibilities for the collective production of scientific information and knowledge. Information Technologies —namely computers connected through Internet— have opened up a field of almost unrestricted possible ways to produce, communicate and share information. Facilities provided by digitalized information for codification, storage, searching, retrieval and combination, as well as the implementation of these task through Internet, make easy, fast and cheap information and knowledge production, communication and dissemination processes. Therefore, it is possible to claim that one of the main functions of these technologies, and, consequently, one of their most valuable features is to ease and increase the creation and communication of information and knowledge. Particular designs and implementations of these technologies should be assessed using these ideas.

However, values of massive information creation and dissemination are not enough to validate these technologies in the epistemological realm. There are good reasons to believe that the increasing amount of information circulating through Internet can result in serious epistemological problems. Information veracity and relevance are the main problems related to the amount of information from any kind of sources available through the computer mediated communication tools. The information accessible on the Internet, at least the amount intended to be trusted as knowledge, can be, and should be, submitted to validation. Information technologies also provide very useful computational tools for that purposes. It is necessary to analyze the new methodologies used for that kind of task and to evaluate them from the point of view of epistemic principles.

Epistemic consequences should be acknowledged an important place within the realm of assessment parameters for these kinds of technologies, as long as they play such an important role on any knowledge related task. New ways of publishing mediated by computers and Internet and new production groups and structures aim at substituting classical institutionalized methodologies (Harnad, 2001). These new ways of producing and disseminating knowledge through Internet have become a topic of ethical, practical and economical discussions. We propose here that these controversies should be regarded, primarily, as epistemological controversies. Therefore,

epistemological concerns about collaborative works, open access publications, open peer reviewing systems, e-prints and self-archiving systems of scholarly publishing will be central to properly frame philosophical discussions. The role of academic institutions and traditional commercial journals should be reformulated in the light of the new computational possibilities.

It is true, however, that the most powerful reasons to claim for an open access knowledge system on the new electronic noosfera are moral reasons. These technologies are, indeed, a main philosophical challenge for information ethics (Feltrero, 2005; Floridi, 2002; Lipinski and Britz, 2000). But moral aspects related to information property systems or knowledge access conditions are also related to epistemological issues. That is why we suggest that epistemological and moral concerns should be articulated (Bustos and Feltrero, en prensa) and that the role and responsibility of researchers, academic and social institutions should be clearly defined to guarantee the proposed articulation.

We deploy here the notion of *epistemic site* as a tool for assessing the epistemological consequences of the use of computers and Information Technologies for the activities of production, communication and distribution of information and knowledge. Knowledge evaluation systems, computer mediated peer reviewing systems and open access initiatives are subject to epistemic evaluation. It is necessary to define and to apply epistemic values and principles for assessing and comparing the epistemic possibilities provided by these new methodologies. Within this framework, we will try to define the conditions that an Internet site has to satisfy for its content to be considered valuable knowledge. In order to ensure the evaluation systems, we propose a new commitment of the most relevant epistemic agents to engage in these new computer mediated systems to produce, select, communicate and evaluate scientific knowledge. The case study of Open Access E-journals is the best way to understand how computers are changing scientific publishing and how we can use epistemic principles, as the notion of *epistemic site*, in order to evaluate this kind of initatives and, eventually suggest new methodologies to improve the epistemic outcomes of the use of computers for publishing.

Open Access publishing through computers

Electronic journals have become popular tools among the scientific community. They are easy tools to store, retrieve, search and manage scientific papers in our own computer avoiding wasting time in the libraries. Computers facilities make possible the easy creation of digital journals making also cheaper most of the editorial tasks. The movement from traditional to electronic journals has led, not surprisingly,

to a big increase in the number of available journals. However it has also led to a big increase of subscription costs (Odlyzko, 1999). Even though commercial editors have reduced their marginal costs, they are increasing the subscription costs making more and more difficult article visibility and quotation impact and, therefore, becoming a hindrance for the collaborative process of knowledge production and assessment.

Open Access is an emerging movement aiming at taking advantage of computational technologies to provide open and free access to research and scholarly publishing. It arises as an alternative to the commercial uses of scientific and scholar publishing. Open Access e-Journals are the canonical example of this movement[224], but there are many new ways in which computers and Internet can help to provide open access to scientific papers. For instance, *Self-Archiving* practices, by the authors or their institutions, are related movements aiming at taking advantage of computational possibilities to store, search and retrieve articles. *E-prints* are the pre-formatted files that are usually offered on the Internet. Many authors also offer their e-prints in their own web page[225]. On the other hand, many universities and institutions are managing repositories with the papers of researchers that belong to them.

The advantages of open access are commonly acknowledged. For instance, Dominy (2006) points out how these electronic media increase the worldwide access to scientific literature, the opportunities for collaboration among experts and the speed to disseminate scientific literature within electronic communities. Others (Harnad and Brody, 2004; Pringle, 2004) show how alternative open access systems as e-prints increase the articles visibility and citation impact. Controversies about open access and peer review systems on the net are, in any case, an old epistemological issue that can be traced back a decade (Fuller, 1995; Harnad, 1995, 1996a). In spite of good reasons supporting open access and the fact that it is an old topic within the academic community, the commercial (toll access) publishing model is still the prevailing model in most research fields. Mainly, surprisingly enough, among researchers in humanities. Being the humanistic disciplines the most “endangered academic species”, researchers do not take advantage of the possibilities Internet offers them to get the maximum visibility and dissemination of their papers and therefore, getting more audience to their social, philosophical or conceptual claims.

The reasons that explain the lack of success or the Open Access Movement are complex and related, mainly within the humanities, to the difficulties of getting researchers out of their traditional methodologies. Economical reasons do not seem to be enough to motivate changes and

224 See http://www.doaj.org/home for a full list of Open Access E-journals

225 Some examples of Self-Archiving are ArXiv (http://es.arxiv.org/), E-prints (http://www.eprints.org/) o Cogprints (http://cogprints.org/)

that is probably because those reasons just point out the advantages of "free" (gratis) access instead of the more particular advantages of "open" access. Even though most these initiatives are completely free for authors, some open access e-journals, far from being free (gratis), can even cause additional costs to them. For instance, some high rank journals, as those belonging to the Public Library of Science initiative[226] are now asking authors for different amounts (up to $2000) to get their papers published. The amount is justified by means of the costs of servers and reviewing systems but can be negotiated, even waived, if authors can prove they belong to countries or academic communities without fundings for research. It is also possible to find hybrid models of toll access and open access. In some cases[227] authors can decide between open access —and in this case the have to pay to get published— or toll access —if they do not want to pay.

Therefore, economical reasons seem to create a more complex publishing world for researchers[228]. How could we get them involved in this movement? Maybe some philosophical reasons could help. Moral reasons seem to be the strongest justification for the open access movement. Justice and fairness demands an equal access to information and knowledge (Lipinski and Britz, 2000) and free access is, for sure, the best way to ensure this equity. Academic justice and fairness demands equal access and also equal possibilities to get published and evaluated. The possibility of a decrease in the quality of journals and academic research due to open access and open publishing tools is a usual argument used to show that obtaining good papers and good knowledge out of these systems is a utopia and how easy it is to get from them more and more useless, and even wrong, information.

The February 2002 Budapest Open Access Initiative[229] declaration and the October 2003, Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[230] are proposals articulate the economical, moral and academic reasons to defend and extend open access e-journals, Those initiatives aim at promoting this movement and getting researchers involved in the creation of a more open and fair system of academic publishing on the Internet. But there is still a lot of work to do in order to ensure that the open access initiative would

226 See http://www.plos.org/journals/

227 See, for instance, The Company of Biologists on http://www.biologists.com/web.openacess.html

228 Not very much complex than the one created by economical concerns on trade publishing. It is an habitual practice on many toll access journals to ask authors to pay in order to get their paper published. Justifications for that are the same that in the open access case (costs of peer review systems) but less proved since commercial journals are supposed to get that money from subscription fees.

229 http://www.soros.org/openaccess/read.shtml

230 http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

contribute to a more open and qualified knowledge production system. Philosophy could help now by means of epistemological reflections.

The epistemological issues related to Open Access are based on the assessment of new on-line possibilities for collective knowledge production through global communication (on-line or asynchronous) and digital processing of digital information. The electronic advantages of the computational communication and processing of scientific knowledge are based on the possibilities to achieve virtually unrestricted storage and publishing within a digital world that, in itself, provides several tools for codification, searching, retrieval and combination of scientific information. But, to argue against the vision of Internet open access sites as just information repositories instead of knowledge resources, it is necessary to point out the advantages that these systems can provide to verify the reliability of those Internet sources. New on-line, open and transparent peer review methodologies can be computationally enhanced in order to prove the epistemic advantages of open access systems. For that purpose is it necessary to elaborate an adequate notion of the epistemic requirements for an Internet site to be acknowledged as a *knowledge* repository, that is to say, what we will call an *epistemic site*.

Epistemic Sites

Among the relevant consequences of process of Information Technologies assessment, those related to knowledge production and communication are central to epistemological concerns. Precisely, those related to truth-values in the activities of belief production and dissemination. These kind of epistemological consequences are analyzed, among other disciplines, by the veristic social epistemology, in the sense defined by A. Goldman (1999). They are subject of assessment in two contexts: epistemic states and epistemic practices. The later can be analyzed in the framework of veristic instrumental values (Goldman, 1999: 87), and that framework will be useful to our purposes. The aim here is to deploy the basis of such framework to assess the computational possibilities for new and valuable epistemic practices in the activities of production and communication of scientific knowledge.

There is a common agreement about the idea that epistemic normativity in science depends upon peer criticism and the fact that acknowledgment is finally acquired by means of the scientific community consensus (Goldman, 2002; Longino, 1990). Community consensus is not the only condition to guarantee true knowledge since it depends on other cognitive and epistemic values, but as the aim of any scientific proposal is to be verified to be considered knowledge, the role of peers is

unavoidable for the whole scientific enterprise. In fact, a researcher can just make a contribution to the common knowledge if her papers are read, cited and can contribute to the development of other researchers papers. The epistemological question now is, are computers and new information technologies changing the ways consensus is achieved in the scientific communities?

The popularity of Internet search engines and their quantitative methodologies to calculate the popularity and impact of web pages can lead us to think this way:

> "I think the Internet community is already naturally telling us what can be considered as "valuable". For example, a site with a lot of external pointers is considered as a valuable site. What is interesting here is that there are no processes, no principles telling people to do so. And, these anchor marks are already used by search engines to determine relevant/trustee sites. (IACAP-2006, anonymous referee)

This kind of argumentation could be interpreted as saying that quantitative consensus is enough to ensure true content. However, traditional and valuable epistemological practices for scientific publishing do not trust on quantitative methodologies. Peer reviewing systems are based on empirical and conceptual discussions about the data and arguments proposed by researchers. Questions and answers are required to improve proposals quality —by means of re-calculating wrong data or re-writting weak arguments— and ensure that contents are relevant for scientific community. Since scientific journals are the preferred methodology to communicate new advances and to submit them to the community evaluation, it is mandatory to apply peer review systems to those journals and to define conditions to evaluate them. That is why, institutions as the European Science Foundation (particularly its section called European Reference Index for the Humanities, ERIH) are demanding the adoption of normal international academic standards to ensure that selection of articles is based on an objective review policy. ERIH is trying to encourage top-journals to adopt a coherent peer-review system. The development of those systems must fulfil epistemic criteria to ensure quality control through the peer review activities and also to open up ways to metaevaluate the very process of peer reviewing.

Nowadays, process of peer review is managed by the journals editorial board that select a few scholars, specialist on the topic, from a pool of volunteers to read and evaluate the paper. Some processes are just a question of marks, whose sum is enough to decide whether the paper should be published or not. More accurate systems ask for comments and arguments to the reviewers that are forwarded to the authors, providing a way to re-elaborate and improve the papers with reviewer's suggestions. Therefore, peer-reviewing systems become part

of the scientific methodology to produce knowledge. Peer review accomplishes several functions. It is a filter to avoid wrong results or pseudo-science to be published, it means the scientific stamp of approval of the contents and methodology deployed by the author and, when it is done properly, is a way to improve communication and collaboration among researchers.

Most acknowledged journals nowadays claim to have the best researchers among their editorial boards and peer reviewers, to justify their claims about the relevance and quality of the articles published on them. But, as this process and its results are not published, there is no way to evaluate how the claimed quality has been achieved. This is a problem since there are many examples that prove the system does not work as well as suggested (McCook, 2006).

The most well known problems are the fake, fraudulent or intentionally bad papers that get published after peer reviewing. The Sokal Affair[231] is one of the most studied cases for humanities and the very last fraudulent paper was the case of the acknowledged magazine *Science* that published a paper by Dr. Hwang about his achievements on stem-cells that was proved to be false (Semir and Revuelta, 2006). On the other side, bad reviewers reject many good papers containing valuable knowledge. These cases indicate that the system is not working properly and should be submitted to evaluation. Indeed the whole scientific community evaluates it, but only after the paper gets published and can be evaluated by a larger community (that usually and finally is able to discover fake results). This point gives us a first clue to understand the steps needed to improve peer reviewing by the use of computational technologies. Maybe those technologies could be used to augment the number of revisions and publishing reviews alongside papers. But for evaluating these new possibilities, we need a new conceptual framework within social epistemology to define how the content can be proved to be knowledge. The concept of epistemic site can be a good starting point for this enterprise.

The notion of *Epistemic Site* aims at deploying ways to epistemically evaluate the content of an Internet site offering knowledge. It can be defined in the following way: *an Epistemic Site is an Internet site offering allegedly true information (alleged knowledge) whose reliability and truthfulness has to be based on, and justified by, a relevant and acknowledged reviewing system*. Of course this is a minimal definition that needs some extra criteria to build upon epistemic relevance and truthfulness with a hierarchy of reviewing systems. The next criteria can help to the purpose of assessment of reviewing systems. *The reviewing*

231 For a brief introduction, see http://en.wikipedia.org/wiki/Sokal_Affair

system has be assessed according to 1) Plurality of mechanisms and referees; 2) Referees' Expertise; and 3) Openness and transparency of the reviewing system criteria and outcomes. Criteria 1 and 2 try to combine the two main advantages of peer review: the possibility to get as much revisions as possible, made by the most expert referees on each topic. Criterion 3 is, in fact, a meta-criterion for criteria 1 and 2 since the transparency is mandatory to evaluate the mechanisms and the referees contributions.

In order to study the open access journals case these criteria will be useful to classify those journals attending to their peer review system, and establish a hierarchy among the different possibilities. The epistemological benefits of open access are first looked for in the visibility and citation impact of freely available articles on the Net (Harnad and Brody, 2004; Pringle, 2004). No doubt, these are epistemological advantages because they increase the knowledge flow, and they are also practical advantages for the scientist since they it increases their possibilities to advance on the traditional academic rewarding systems based on the number of citations. Not surprisingly these are main criteria for assessing journals on the international indexes as ISI (Institute for Scientific Information). We claim here that it is necessary to evaluate the epistemological benefits of computational technologies applied to publishing with qualitative methodologies. The benefits of visibility and citation impact belong to quantitative methodologies, and the notion of epistemic site can help us to prove that open access is a qualitative epistemic advantage that could lead to more accurate and epistemically valuable electronically enhance peer review systems.

At first sight, it is possible to find new alternatives on peer review in open access e-journals. Since many of them have a traditional methodology of previous peer review and article selection, (that is the case of, for instance, the Public Library of Science, http://www.plos.org) it is possible to find two alternative methodologies. The first one is the absence of a previous peer review system as it happens in most of the self-archiving systems (ArXiv is a good example, http://www.arxiv.org). In these cases, authors just use the system to store their paper and to open it to the evaluation and criticism of peer researchers. The peer review system is based on parallel e-discussion groups. Most self-archiving systems provide the author with the possibility to store the following versions of their corrected papers. Finally, it is possible to find an increasing number of open access e-journals (and also toll access e-journals) that offer several possibilities to make on-line Peer Review. Usually, peer commentaries are published beside the articles and, many times, the commentaries can be added by everyone by means of e-letters or editing tools directly on the Internet site.

Are open peer commentaries epistemically equivalent to an open peer reviewing? For Stevan Harnad they are not equivalent as long as he clearly distinguishes between the roles of peer review and peer commentary (Harnad, 1998). He claims that open peer commentaries should be a supplement, not a substitute, for peer review (op. cit. p. 1045). This is based in a division of peer review roles. On one hand, he sees peer review as a pre-publication selecting system and open commentary as a post-publication peer commentary. We consider this vision misleading. Peer review and peer commentary can be the same activity when publishing on computational technologies. Computers provide us with the possibility to publish the successive revised versions of one paper. So, regardless whether the paper has been published or not, peer commentaries and reviews are contributing to knowledge production and quality.

Among traditional journals, revision is first and commentary comes after publication. Technology can reverse the usual process in order to get, first, massive open peer review/commentary capable of contributing to ensure the truthfulness, veracity and accuracy of the knowledge contribution and, at the same time, increasing it; and, second, an expert peer review system can be impliemented to assess and promote the most relevant topics and discussions held in an open access basis. Given this epistemological definition and insights, the question now is how computational technologies can implement these epistemological criteria and how computational publishing can enhance and improve in epistemological relevant ways traditional peer reviewing systems.

Electronically enhanced peer review systems

Interactivity and multimedia are the most relevant new features of scientific communication mediated by computational technologies. Interactivity speeds up traditional ways to communicate by means of e-mail, distribution lists or chats. Multimedia allows us to exchange not only text, but also any kind of data and scientific representations related to the content of the knowledge piece that is being communicated.

Several mechanisms of interactivity have been widely studied by Stevan Harnad (Harnad, 1995, 1996a, 1996b) giving rise to his concept of “scholar skywriting”. Hovewer, these systems are not yet widely used for peer reviewing activities. And there are many new choices. An special place has to be reserved for the browsing/editing tools in Internet that allow us to directly participate on collaborative web pages (the free encyclopaedia *Wikipedia* is the best example of these tools possibilities). Wiki-like tools are capable of having a big potential to contain on-line

open peer review. *Blogs* are good tools for authors to keep communicated with their readers providing post-publication comments. But, for sure, there are many ways to electronically articulate peer review systems according to our epistemic criteria. Some of them can be found on attempts to design Cooperative and Collaborative Computer Supported systems (Sumner and Buckingham Shum, 1998; Sumner, Buckingham Shum, Wright et al., 2000). The JIME project[232] is a very good example on how computational technologies can contribute to change the current academic publishing practices. From traditional peer review systems deployed as anonymous, mediated and almost monologic "vetting" process, collaborative tools can lead to a new system where peer reviewing is conceived as a constructive design process to improve both author and reviewers knowledge and contributions. Computational technologies provide tools to get every participant, that is to say, authors, reviewers, editors and finally readers, engaged in a direct dialog capable of promoting more dynamical and enriching publishing systems (Sumner et al., 2000, p. 5).

A good electronically enhanced open peer review system has to provide a transparent peer review system (under the principle of open access to every information related to the paper and its reviews) capable of making explicit arguments and scientific controversies. And has also to provide transparent outcomes for other authors to build upon. This last function aims at merging knowledge production and evaluation systems. The former can contribute to bridge the gap between science production and science communication, helping the interested people to acquire the scientific knowledge relevant to their own social or personal context and to participate in virtually unrestricted public discussion about science and its aims.

These methodologies can lead to produce reliable and truthful e-journals with a dynamic intellectual production, and can also help to implement some automatic accounting methodologies by means of the number and sources of received reviews and triggered controversies[233]. Technological aspects are being solved as time goes by, but the very revolution can only be real if we get experts involved in the open peer review world. For that purpose, academic institutions have to design a way to formally reward peer review activities. In this case, open access e-journals can also ease electronic account of peer reviewing production in order to get an adequate account of these rewarded activities.

232 http://www-jime.open.ac.uk/

233 Citesser tools (http://citeseer.ist.psu.edu/cs) can mean a first sep towards this kind of automatic systems, but still to much quantitative.

Conclusion

Computational technologies provide tools to revolutionize the traditional ways of scientific publishing based on asynchronous print-paper technologies. This can be done by means of interactive —synchronous or asynchronous but faster— methodologies through computers connected to the Internet. These methodologies could ease open, universal and free knowledge access for everybody, which is indeed a moral target. The Open Access movement is driven towards this direction and the increasingly number of open access e-journals, with an increasing impact on the epistemic communities, means a revolution. But this revolution is still incomplete. Computational technologies also provide tools to deploy open peer review and peer commentary systems and the next step towards this revolution should be to implement these kind of systems on the new open access e-journals and to get experts to participate on those reviewing activities.

The notion of epistemic site implies that a quality peer review methodology should be the guarantee for an Internet site to prove it contains knowledg. This is also valid for scientific journals (and particularly e-journals) to get acknowledged a as "good" journal. In order to ensure this, that methodology has to be open and inspectionable somehow, that is why interactive open systems are more suitable to get this quality. From this point of view, free access is not mandatory just for moral reasons, but also as a way to improve the epistemic quality of scientific publishing. Anyway, the recollection and organization of "peer resources" (even by paying them) to get experts involved in a good peer review system could be a commercial activity, managed by traditional trade journals on the basis of open documents and open peer review systems. Free access still leaves open alternative ways to get money and influence, and this ways could still be a good source of benefits for commercial publishing houses, as it happens with very profitable business based on the Net. Anyway, the more open and epistemically fruitful way to improve peer-reviewing systems with expertise is to stimulate experts participation by means of a system of academic rewarding. Being peer review a way both for intellectual and academic improvement, the system should work under their own epistemic constraints.

Open Access digital on-line papers could also be the most appropriate framework to deploy scientific publishing into a more collaborative, transparent, fair and socially accessible activity. Within these advantages, we have advocated here that the epistemic ones should be remarked in order to extend the open access evolution provided by computers and Internet to the strong revolution this

technologies can cause by means of the implementation of open peer review methodologies. It has been shown that an electronically enhance peer review system can provide multiple assessment methodologies for every scientific article where evaluation is driven, first, by epistemological concerns and, in a second step, by the magazine assessment and classification. This could lead to a more democratic assessment (on principle) without the risks or anonymous editorial mediated peer review system. Once the first step of open peer review systems is taken, plenty of new social methodologies to produce scientific and academic publishing could be imagined and implemented. Social epistemology, in this way, can be complemented by a branch of "methodology design" of this new computational mediated ways to produce and evaluate knowledge.

To get fully electronical and open systems, further steps are needed. A standard document format in some marked language (as XML) will help to implement data mining through papers. With this kind of marked e-prints, not only it its not necessary to add metadata files with the authorship, keywords or abstracts, but also can be collected, references, quotations, commentaries, referees, etc., in order to directly and automatically harvesting every datum relevant to make every kind of epistemic evaluation of the paper impact on the relevant intellectual communities. Finally, a clearer and more standardized open licence system for academic publishing should be put into work to preserve and guarantee, at least, moral and intellectual (copy)rights for authors and researching activities. Again, apart from technical, legal, economical or moral arguments, all these further steps can be justified, primarily, as a way to implement valuable epistemic practices.

Acknowledgements

This work was funded by the Spanish Ministry of Science and Tecnology research project HUM2005-02105/FISO.

Bibliography

Bustos, E. y Feltrero, R. (en prensa). Internet y el Acceso al Conocimiento: la articulación de lo Epistemológico y lo Moral. In A. Pérez y A. Velasco (Eds.), *Racionalidad teórica y racionalidad práctica en la ciencia*. Mexico: Facultad de Filosofía y Letras, UNAM.

Dominy, P. y Bhatt, J. (2006). *Peer Review in the Google Age. Is technology changing the way science is done and evaluated?* Available [Last Access: 2006, June].

Feltrero, R. (2005). *Ética y Valores en la Sociedad Informacional*. Madrid: UNED.

Floridi, L. (2002). On the intrinsic value of information objects and the infosphere. *Ethics and information technology,* **4**, 287-304

Fuller, S. (1995). CyberPlatonism: an Inadequate Constitution for the Republic of Science. *Information Society,* **11**(4)

Goldman, A. (2002). *Pathways to Knowledge*. New York: Oxford Univ. Press.

Harnad, S. (1995). Sorting the esoterica from the exoterica: there is plenty of room in Cyberspace: Response to Fuller. *Information Society,* **11**(4), 305-324
http://cogprints.soton.ac.uk/documents/disk0/00/00/16/84/index.html

Harnad, S. (1996a). Implementing Peer Review on the Net: Scientific Quality Control in Schorarly Electronic Jounals. In R. Peek y G. Newby (Eds.), *Scholarly Publishing: The Electronic Frontier*. Cambridge MA: MIT Press.

Harnad, S. (1996b). Interactive Cognition: Exploring the Potential of Electronic Quote/Commenting. In B. Gorayska y J. Mey (Eds.), *Cognitive Technology. In Search of a Humane Inteface.* Oxford: Elsevier.

Harnad, S. (1998). Learned inquiry and the net: The role of peer review, peer commentary and copyright. *Antiquity,* **71**, 1042-1048

Harnad, S. y Brody, T. (2004). Comparing the impact of Open Access (OA) vs. Non-OA articles in the same journals. *D-Lib Magazine,* **10**(6)

Lipinski, T. A. y Britz, J. J. (2000). Rethinking the ownership of information in the 21st century: Ethical implications. *Ethics and information technology,* **2**(1), 49-71

Longino, H. (1990). *Science as Social Knowledge*. Princenton: Princenton Univ. Press.

McCook, A. (2006). Is Peer Review Broken? *The Scientist,* **20**(2), 26
http://www.the-scientist.com/2006/2/1/26/1/

Odlyzko, A. (1999). The economics of electronic journals, *Technology and Scholarly Communication* (pp. 380-393): Univ. Calif. Press.

Pringle, J. (2004). Do open access journals have impact? *Nature* (Web Focus) http://www.nature.com/nature/focus/accessdebate/19.html

Semir, V. y Revuelta, G. (2006). Dr. Hwang and the clone that never was. *Quark,* **37-38**, 105-123

Sumner, T. y Buckingham Shum, S. (1998). From Documents to Discourse: Shifting Conceptions of Scholarly Publishing., *Proc. CHI 98: Human Factors in Computing Systems* (pp. 95-102). Los Angeles, CA: ACM Press: NY.

Sumner, T., Buckingham Shum, S., Wright, M., Bonnardel, N., et al. (2000). *Redesigning the Peer Review Process: A Developmental Theory-in-Action.* Presented at the COOP'2000: Fourth International Conference on the Design of Cooperative Systems, Sophia Antipolis, France.

THE BEHAVIORAL APPROACH IN COMPUTER SIMULATION AND ROBOTICS

JOHANNES LENHARD

Department of Philosophy

Bielefeld University

Phone: +49 521 106 4597

Email: johannes.lenhard@uni-bielefeld.de

Introduction

Computer simulation originated around 60 years ago and has evolved into a widespread method of science and technology. The present paper will identify two early and very influential paradigms of simulation modeling – one in conflict with the other. The first is associated with the name of John von Neumann. It conceives of simulation as the *numerical solution of equations* – drawing on the qualities of the computer as number cruncher. The second approach, elaborated most prominently by Norbert Wiener, views simulation as *imitation of the*

behavior of a complex system by using a computer model – underlining how versatile computer systems can adapt to patterns.

Wiener and von Neumann planned to combine their skills with those of other colleagues in an interdisciplinary group aiming to explore the potentials of the newly developed computer as an instrument for science. Both researchers were extremely influential in the early phase of the so-called *Cybernetics Group*. However, very early on they began to have serious trouble with the conflicting modeling conceptions they held. Today, more than half a century later, simulation methodology has become a widely used instrument. More precisely, it will be argued, it is based on an amalgamation of the two conflicting approaches to modeling. Simulation methodology decoupled the imitation of behavioral patterns from the approximation of a solution. In other words, simulation models can imitate phenomena without having determined the laws of model behavior that underlie them. That analysis applies for a wide range of simulation modeling techniques.

Instead of going into the details of that claim, the second part of the present paper will focus on the so-called behavioral approach in robotics that claims to present a new conception of artificial intelligence (AI), in particular a more successful one than traditional AI. The analysis of the conflicting conceptions of simulation modeling from the first section will be used to shed some light on this new approach in robotics and AI. The argumentation will concentrate on Rodney Brooks, MIT, and his conception of behavioral robotics.

Some important continuities will be marked, most importantly the criteria of modeling success. However, the new behavioral approach differs significantly from its forerunner. Above all, it is claimed, the new approach in robotics combines traits of both basic modeling conceptions and hence represents a kind of hybrid. While the behavioral stance is located in the camp of "weak" AI, some claims of behavioral robotics involve the correspondence to the real way biological systems bring about intelligent behavior. Hence one can speak of a strong behavioral approach. Thus the established coordinate system of weak vs. strong AI that gave rise to so many quarrels seems to be transformed.

This transformation, it is argued, is based on a revaluation and re-interpretation of AI that is deeply influenced by the available instruments, i.e. the technological development of the computer and computer systems. This revaluation, it will be concluded, may be interpreted as part of an ongoing transformation in scientific culture, triggered by the interaction between computer, philosophy, and science.

Two Types of Modeling

Major components and principles of the electronic computer were worked out during the 1940s. The development and implementation of various analogue and digital calculators took place mainly within the framework of war-related military research. The beginnings of computer simulation, in the following conceived as digital technology, form a part of this history with the mathematicians Norbert Wiener and John von Neumann playing a major role as influential "founding fathers." Both had the vision that the new computer technology combined with corresponding new approaches to mathematical modeling would lead to an epochal reform. They planned to combine their skills with those of other colleagues in an interdisciplinary group aiming to explore the potentials of the newly developed computer as an instrument for science. This was the context for the founding of the *Teleological Society* that led to the *Cybernetics Group*, which first met secretly in January 1945 (in Princeton), then officially from 1946 on in New York in ten further 'Macy conferences'.

Wiener and von Neumann both were active in wartime research. Indeed, the wartime situation and the corresponding pressure to produce applications exerted an enormous influence on mathematical theory formulation (see, for instance, Heims 1980 or Galison 1994). Wiener, who worked at MIT and was very closely linked to engineering, worked with the engineer Bigelow, commissioned by the National Defense Research Committee (NDRC), on an Anti-Aircraft-Predictor, a computer-based defensive system. The NDRC project also marked the beginning of Wiener's lasting interest in the links between regular processes and goal-directed behavior. He assigned great philosophical significance to these, as recounted in the programmatic paper "Behavior, Purpose, and Teleology" (Wiener, Rosenblueth, and Bigelow 1945). Through the synthesis of the new technologies with his theories, Wiener envisaged the onset of a new epoch, the cybernetic age (Wiener 1948).

John von Neumann, a mathematician at the Institute for Advanced Studies in Princeton, also foresaw the fundamental significance of the new computer technology. He was involved not only in its further development, particularly of the architecture for the general purpose computer named after him, but also in the invention of new simulation approaches (see, e.g., Galison 1996, on the origins of the Monte Carlo simulation in the context of the Manhattan Project). Put briefly, both protagonists were interested in developing a new scientific discipline that would advance the computer to a new general-purpose instrument.

Despite all they had in common, Wiener and von Neumann held conflicting conceptions of modeling and simulation. During the planning

phase of the *Cybernetics Group*, a rupture began to emerge due to different conceptions of the terms 'imitation', 'understanding', and 'modeling'. Wiener developed cybernetics as a science that examines phenomena and models according to their functionality and behavior, and not according to their material and inner structure. Ashby emphasizes this in his introduction to cybernetics (oriented toward Wiener):

"Cybernetics … does not ask "what *is* this thing?" but "*what does it do*?" Thus it is very interested in such a statement as "this variable is undergoing a simple harmonic oscillation," and is much less concerned with whether the variable is the position of a point on a wheel, or a potential in an electric circuit. It is thus essentially functional and behaviouristic." (Ashby 1957, 1)

Wiener's approach which viewed behaviorism and feedback as one philosophical unit and treated, in particular, human beings and machines in a completely analogue way was subject to controversial discussion. In the second part of the present paper I will discuss how the behavioral approach in robotics ties in with this line of cybernetics. Moreover, Wiener based his approach on a functionalistic concept of models. In an article with Rosenblueth, he found that models are indispensable and that science aims simultaneously at control as well as understanding , implying a "dualistic attitude" (Rosenblueth and Wiener 1945, 316).

There is namely no guarantee that the efforts to control will also lead to understanding or vice versa. As a result of this, they distinguished between *open box* and *closed box* (or, as more commonly called today, *black box*) approaches. This terminology was adopted from a communication technology using test procedures to evaluate an instrument according to input-output patterns regardless of the mechanism within. Although these types of boxes differ only gradually, this difference describes a typology imposed by the complexity of the applications: *Open boxes* may be a fine ideal, but not one that can be used in an applied orientation – and the behavioral approach that measures modeling success by the imitation of behavior patterns does not just acknowledge this but fundamentally has no alternative.

Hence, there are two competing types of model: The *open boxes*, which contain a more or less detailed translation of the structure and represent laws or mechanisms versus the *black box* models that can only be treated in terms of their behavior (i.e., functionally) and imitate behavioral patterns without making statements on the internal dynamics of the phenomena being modeled. Wiener considers black box modeling to be necessary for not only philosophical (uncertainty) but also pragmatic (applicable technology) reasons. Humphreys (2004), in his recent systematic account of simulation, brought up a similar issue and

pointed to *epistemic opacity* as one, nonetheless deplorable, feature of simulations.

John von Neumann associated a completely different modeling strategy with simulation. He adopted a far more formal, one could even say more optimistic stance, that viewed the computer as an aid in mathematical theory formulation, and engineering applications as providing new ideas for theoretical development. Most importantly, he rejects Wiener's program of cybernetics almost entirely, insisting instead on modeling the right mechanisms. In von Neumann's view, a Wienerian approach is a bad modeling strategy precisely *because* of the performance and adjustability of simulation models: It is impossible to learn anything about the right mechanisms from a successful imitation of behavior patterns ("functioning"). The path from a functional to a structural modeling would be cut off; von Neumann sees that completely correctly. He makes the possibility of structural models, admittedly without realizing the principal contradiction to Wiener, the criterion of research planning (cf. his long programmatic letter to Wiener from 1946, printed in Masani 1990).

Von Neumann saw the usefulness of computer simulation mainly in fields where fundamental equations exist, but cannot be solved for reasons of complexity, in the sense of missing computational power. His work with Ulam on Monte Carlo integration during the Manhattan project represents an example. Particular mention should be given here to his programmatic approaches to a numerical solution of hydrodynamic equations. He was convinced that systems of partial differential equations could be tackled numerically. In this way, the computational power is the key for simulation methods – making it possible to solve problems where other modeling approaches are inadequate. Consequently, von Neumann initiated a group at Princeton that tried to tackle meteorological problems, namely, solving the fundamental equations of the general circulation of the atmosphere.

The story of Wiener versus von Neumann could be elaborated to a full-fledged case study on its own, comprising the success of von Neumann's group at Princeton and climate simulations, as well as Wiener's accusations for philosophical reasons. Küppers and Lenhard (2005) show that the achievements of Neumann's simulation strategy, now hegemonial in climate research, owe much to the inclusion of a Wienerian (or say behavioral) strategy. It should be noted that the behavior of the simulation models is important in both approaches. In von Neumann's sense, it is used as a criterion for whether the mechanisms and laws have been transformed adequately, whereas "behaviorist" implies that the concern is this behavior itself and not what goes on inside the black box.

His overall goal was "true understanding," and that requires working with structurally isomorphic mechanisms. He rejects the functional concept of modeling in favor of the structural one. He takes the versatility of functional modeling as an exclusion criterion whereas the imitation ability was the basis for Wiener's high-flying hopes for cybernetics! Basically, both succeeded in pushing through their research agendas: Wiener founded cybernetics in his sense, while von Neumann left the Cybernetics group, carrying on his projects independently.

The controversy between von Neumann and Wiener is of more than historical interest, because it exemplifies the fundamental conflict between the modeling approaches. This can be traced back to the quarrels between Descartes and Galilei, or Leibniz and Newton, and in recent simulation methodology it breaks up again in a transformed way. Of course, the controversy between the two modeling approaches is widespread and not limited to meteorology. Von Neumann himself introduced the virus, because he thought it would be possible to model its mechanisms. On the other side, Wiener was preoccupied (along with McCulloch, Pitts, and others) with human neurophysiology. In what sense can the computer simulate intelligence? Right from the beginning, the artificial intelligence (AI) community was divided on this issue. The behavioral approach had already been taken by Turing (who, by the way, was very familiar with Wiener), who conceived his test expressly as an "imitation game." This kind of black box approach, taking only input-output behavior as the criterion, is also known as "weak" AI. This is contrasted with so-called "strong" AI that insists on using only models that implement the correct mechanisms.

The behavioral approach in robotics

Let us move into the present time and discuss today's celebrated behavioral approach in AI. The previous analysis has equipped us with conflicting standpoints in simulation modeling that will serve as a point of reference. To be more precise, it will be claimed that this new approach combines traits of both sides of the conflict, thereby questioning the established coordinates of strong vs. weak approaches in AI.

Since the 1990s a few researchers in artificial intelligence (AI) have come to the conclusion that their own discipline had produced results alarmingly poor in relation to the great expectations they had had. They argued that a new and fundamentally different approach would be necessary in order to achieve significant progress. Researchers like Rodney Brooks (2002), Rolf Pfeifer (1999), or Luc Steels (1995) among

others promoted a new conception of AI research. Naturally, they stressed different aspects and hence the new approach did not come up as a coherent program. Nevertheless, the concepts of situatedness, embodiment, and behavior present common coordinate axes of the different proposals – inviting one to treat them as one (albeit heterogeneous) movement. In the present paper, argumentation needs to be restricted, hence it will concentrate on the "behavioral approach" as has been suggested by Brooks. Other closely related views cannot be discussed for reasons of brevity; I hold that such a discussion would strengthen the claims of the paper. Brooks contrasts his approach to the "traditional" one:

„Traditional Artificial Intelligence has tried to tackle the problem of building artificially intelligent systems from the top down. It tackled intelligence through the notions of *thought* and *reason*. (…) Recently there has been a movement to study intelligence from the bottom up, concentrating on physical systems (e.g., mobile robots), situated in the world, autonomously carrying out tasks of various sorts. (…) The flavor of this work is quite different from that of traditional Artificial Intelligence." (Brooks 1991, 1)

How does he spell out the specific differences? One of Brooks' points is that AI should take the route via robotics (others say: artificial life), hence behavioral robotics will be used synonymously with behavioral AI. It is presented as a modest withdrawal of all too detached prospects for AI: First, build robots that can carry out some real world tasks, before taking the next step. The deliberate goal not to confine the environment to an artificial laboratory setting will require intelligent robots, even for basic tasks like orienting and moving in a room, the reasoning goes. Brooks mentions a second reason for his approach: "In fact it suggests that despite our best introspections, traditional Artificial Intelligence offers solutions to intelligence which bear almost no resemblance at all to how biological systems work." (Brooks 1991, 1)

This second aspect is of crucial importance in characterizing the whole endeavor. Of course, nobody claims that a one-to-one emulation of biological systems would be neither feasible nor favorable at all. Nevertheless, the way biological systems really work plays an important role for the behavioral approach.

First, it claims to follow a pragmatic and modest line, aiming at the imitation of real world behavior, avoiding strong claims about how intelligence functions. Second, it criticizes traditional AI for not simulating biological systems, suggesting that AI should follow the path of actual biological evolution. This, in turn, is a very strong claim. It will be argued that this modest and at the same time strong characteristic relies on a transformation of the philosophical coordinates in which AI is commonly

interpreted, namely what is called 'weak' and 'strong' AI. In short, the behavioral approach combines weak and strong, i.e. antagonistic, traits. The comparison with the conflicting modeling strategies of Wiener and von Neumann will help complete the picture of what constitutes the behavioral approach.

Obviously, it stands in stark contrast to the von Neumann approach. Brooks even states explicitly that the von Neumann architecture has deeply influenced and, moreover, misled the development of AI by suggesting that modeling approaches should include centralized coordination and a general internal representation of the world 'outside'. The behavioral approach tries to overcome this way of thinking about reason that has been imposed by the instrumental basis of early computer development. Brooks calls this "bottom up": tinkering together different, autonomously functioning modules – without central coordination, without general internal representational or world-model. This is nicely captured by the title of his paper "Intelligence Without Reason" (1991).

This stance looks Wienerian: According to black box modeling strategies, the behavior of a model provides the crucial criterion, because the inner mechanisms are not accessible. While Wiener's conception can also be called behavioral (damping down the psychological connotations of the term "behavioristic"), there is a crucial difference between the older and newer behavioral approaches. Brooks not only declares the bottom-up approach to be practically necessary, as a way to escape the unfruitful situation of AI, but he also defends it on a higher level: it is *the right* strategy, precisely because it imitates biology in how it really produces intelligent behavior. Brooks suggests that imitation of behavior is a *strong* criterion as it rules out features that are misleadingly induced by the orientation toward representation.

Thus Brooks advocates an optimistic behavioral approach aiming to reach the top level from bottom up, i.e. a behavioral approach that is claimed to match the real mechanisms going on in biological evolution. Wiener, in contrast, has seen a fundamental incompatibility between real world phenomena and their scientific analysis. For him, behavioral approaches expressed a fundamental pessimism. To accept behavior as a criterion is a kind of unavoidable evil. Wiener held that the underlying mechanisms are inaccessible and that this marks the principal condition of human epistemology. Hence the behavioral approach in robotics represents a remarkable re-adjustment, or rather synthesis of the conflicting modeling strategies examined in the first half of the present paper. How does Brooks himself link his variant with the history of AI and, in a broader sense, with the history of computer and simulation modeling?

He refers positively to Turing and the famous imitation criterion. The point is, according to Brooks and the behavioral approach in general, that to satisfy this criterion, i.e. to produce or simulate intelligent behavior, one should work with behavior-based machines instead of involving general and universal models. Shall we take this approach simply as a preference for an engineering perspective that is comfortable with tinkering together some technological modules and which seeks to convince the community that more theoretical approaches are misleading? I think the story is more complicated– and therefore more interesting as well.

Brooks acknowledges, alluding to the works of Ross Ashby and Grey Walter, that cybernetics had already employed situatedness and showed little interest in representing models. According to Brooks, the crucial difference between behavioral robotics and cybernetics lies in the conditions the available computational instruments imposed. These conditions have changed to a great, some even say incredible, extent. Early proponents such as Ashby, Walter, or Wiener were guided (or, rather misguided) by their experiences with technology in its infancy: „But in some deep sense Wiener did not see the flexibility of these machines". (Brooks 1991, 7)

This marks an important point: Wiener's account of modeling was based on the versatility of the computer to imitate behavioral patterns. The actual development of this instrument, however, has outdated Wiener's views and has shown that the advanced computer and computational models are much more flexible than anticipated. While Wiener thought that fundamental mathematical and statistical theories were absolutely necessary to build black box models, Brooks is much more confident in the potential progress one could make by tinkering together various modules from the 'bottom up'. Hence the actual development and refinement of the instrument, as well as experiences in application, exert a fundamental influence on the epistemology of computer and simulation approaches. So, is the behavioral approach in robotics to be interpreted as a reformulation of older cybernetical views on the basis of half a century's technical elaborations of the computer? It would be premature to interpret Brooks' optimism as resulting from Wiener's pessimism which had been disabused by technological advances.

"What is the point of all this? The traditional Artificial Intelligence model of representation and organization along centralized lines is not how people are built."(Brooks 1991, 13) It is a fundamental philosophical assumption that the behavioral approach presents the right modeling approach in a strong – one could say Neumannian – sense. This assumption is based on a radical revaluation: central parts of weak AI – restriction to imitation of behavioral patterns as feasible criterion – now

acquire a strong interpretation, namely correspondence to reality, i.e. real biological systems.

"Real biological systems are not rational agents that take inputs, compute logically, and produce outputs. They are a mess of many mechanisms working in various ways, out of which emerges the behavior that we observe and rationalize."(Brooks 1991, 14)

How does Brooks combine (weak) behavioral criteria and (strong) claims about how real systems work? Do technological advances lead to a reconciliation of the conflicting modeling types? The quotation shows that the glue is provided by the famous concept of emergence. It bridges the gap between the conviction that *in principle* intelligent behavior is produced by tinkering together different autonomous components, whereas *in practice* the criteria of successful modeling are those of black box modeling.

Indeed, Brooks grants emergence a central place in his conception: There are real mechanisms, but they work together in a messy way that cannot be analyzed in detail for reasons of complexity. Hence the behavior they produce has to be taken as basic constituent – which is the stance of black box modeling. In a sense, Brooks has accepted the Wienerian pessimism: it makes no sense to analyze the real mechanisms biological systems employ to bring about their behavior. At the same time, behavioral robotics has adopted the Neumannian optimism, but on a different level: it is not the real mechanisms themselves that are modeled, but rather *the way* real biological systems acquire behavior. Hence it can be summarized that the optimism of behavioral robotics to achieve progress where the traditional approach failed is based on an analogy of method, not on new insight into mechanism. On the contrary, the concept of emergence acknowledges explicitly that the real mechanisms and their way of interacting remain epistemically opaque – a messy interaction.

How emergence can be controlled effectively, i.e. how the behavioral approach can become a basis for reliable technology, remains a dark spot – the only argument seems to be that tinkering has been proved to be very powerful. The problems arising here have their analogy in simulation modeling quite generally. The trust in the flexibility of computational modules serves as a guideline. Admittedly, flexibility and versatility have been increased to a surprisingly high level. Often this point is mentioned in criticism of simulation modeling approaches: computer simulations bring about, it is said, their behavior by artificial mechanisms, hence simulation phenomena possess no trustworthy relation to their real counterparts. It is exactly this vague, flexible, and elusive relation that is the cornerstone of the behavioral approach in robotics.

Concluding Remarks

What is then to be said about the pretension of control and understanding, the dual goals of science as Wiener and Brooks agree? Brooks is quite aggressive on that point:

„I think that the new approach can be extended to cover the whole story, both with regards to building intelligent systems and to understanding human intelligence—the two principal goals identified for Artificial Intelligence at the beginning of the paper.

Whether I am right or not is an empirical question."(Brooks 1991, 17)

Having the argumentation of the present paper in mind, one could express some restrictions. Control may turn out to be difficult to achieve, in particular when many mechanisms interact. Brooks seems to allude more to a hope than to an argument. On the other hand, his optimism is fed by the fact that engineering approaches have accomplished control in situations where theory did not (or not yet) provide practical approaches—the steam engine provides a standard example here.

The second concept, understanding, raises even more doubts: Doesn't black box modeling *per se* preclude understanding? The refinement of instruments, or better: the emancipation of the computer as a scientific instrument, contributes to the prospects of the behavioral approach. The instrument has become so common that this approach can lay claim to being 'bottom up'. The behavioral approach à la Brooks is actually a 'strong' approach in AI synthesizing the two conflicting modeling conceptions named by and ascribed to Wiener and von Neumann. Whether this synthesis will actually be successful remains an open and empirical question.

It might well turn out that the standards of what counts as scientific understanding adapt to the shifting technological environment: If the behavioral approach is successful in building intelligent robots, then presumably the concept of understanding will be adjusted so that it describes what AI achieved – even if there remains a black box of properties emerging from a messy interaction of modules. If this approach is successful – will we gain a better understanding of intelligence? Due to the changes the concept of understanding may undergo in the course of these developments, this is an open and empirical question as well.

Bibliography

Ashby, W. R. (1957). An Introduction to Cybernetics, 2nd impression. London: Chapman & Hall.
Brooks, R. (1991). "Intelligence Without Reason." A. I. memo 1293, MIT 1293.
Brooks, R. (2002). Flesh and Machines. New York, Pantheon Books.
Galison, P. (1996). Computer Simulations and the Trading Zone. The Disunity of Science: Boundaries, Contexts, and Power. P. Galison and D. J. Stump. Stanford, CA: Stanford University Press**:** 118-157.
Masani, P. R. (1990). Norbert Wiener 1894-1964. Basel, Boston, Berlin: Birkhäuser.
Pfeifer, R. and C. Scheier (1999). Understanding Intelligence. Cambridge, MA
London, England, The MIT Press.
Steels, L. and R. Brooks, Eds. (1995). The Artificial Life Route to Artificial Intelligence : Building Embodied, Situated Agents. Hillsdale, NY, Lawrence Erlbaum.
Wiener, N. (1948). Cybernetics or Communication and control and the Animal and the Machine. Cambridge, MA: MIT Press.
Wiener, N., A. Rosenblueth, and Bigelow, J. (1945). "Behavior, Purpose, and Teleology." Philosophy of Science 10: 18-24.

UNIQUE LOGIC OF THOUGHT

STANISLAW ZADROZNY
USA (505) 897-7222

stanzadrozny@yahoo.com

Introduction

The present world is very different from the world of previous centuries. As never before in history, human activity today depends upon the efficient flow of information from newspapers, radio, TV, and, most importantly, from Internet access by PC users. Although people are not particularly conscious of it, thinking could not exist without information. Thinking is a process done individually, but it could not have developed without intelligent social interaction. It is compounded by the contributions of other people who create, communicate, and gather the results of perceived and cognizant reality. Thinking is impossible without socially evolved languages, numbers describing quantities, and graphs or pictures delineating qualities, all of which allows information to be distributed throughout society.

Our thesis, briefly stated, is that, by utilizing the concept of spaciousness, which may later be simulated by the computer, one will be able to imitate, replace, and extend the process of thinking. Through language, the results of thinking are presented in a linear fashion, while the entire process occurs spaciously within the mind utilizing information as a medium. The premise underlying our thesis is that the construction of a spacious model of the mind on the proper level will enable a transfer of the thought process to a computer, which, in turn, would provide meaningful results.

Actual progress in the field of computer science and information technology has left the philosophical concept of information in its wake. It lacks a clear description of what is or is not information per se. On the side of computer science, there are mega, giga, and tera bytes, along

with everything that can be done by them; on the side of the theory of knowledge, however, a conceptual apparatus that would allow for the full utilization of such potential does not yet exist.

The paramount concern of information technology (IT) is to provide speed, accuracy, and flow of the volume of data being reworked. The theory of knowledge should step down from its ivory tower and essentially provide keys to the interpretation and reinterpretation of these data. It can be assumed that closing the gap between computer science and the theory of knowledge will open up a new path to artificial thinking. Likewise, there is a strong belief that replacing language with information in the basic logical functions is the first step in this direction, since information**,** not language, is the medium through which thinking is realized. In order to specify the conditions under which thinking and its attributes are to be realized, our first question should be: What is information?

Information

The following definition is proposed: **Information is a representation of 'reality' objectified in a given code system. Representation** may be described by synonyms, such as, presentation, reflection, projection, description, depiction, imitation, etc. The meaning of all of these words is similar, and it refers to the relationship between an object and its representation. With the word '**reality',** one understands everything that can become the object of perception, inquiry, cognition, reasoning, investigation, experimentation, presentation, etc. '**Reality**' refers here to physical objects, processes, and also to abstract ideas, notions, concepts, etc. **Objectified** means presented according to commonly accepted rules of operation and supported by a verified method and a proper technique. With the expression **given code system,** one understands a particular form of signs, symbols, and graphs selected for a presentation of the shape of information on a given carrier. An item that represents a single distinguished part of reality will be defined as an element of information (ei).

Levels of Analysis (Vertical Projection)

The acceptance of the concept of an element of information (ei) enables one to view particular information in the abstract. However, the result of representation, i.e., an element of information (ei), also has a vertical projection that can be examined on four levels, which we will call **levels of analysis.**

The level of analysis that describes an element of information (ei) within the framework of particular information is marked as 0 (zero). This represents the relationship between reality and information, as well as categories of cognition; and it also demonstrates that information, when it is free from particulars has a universal character. The zero (0) level of elements of information has a virtual character. The existence of this level implies a necessity for an interpretation requiring the possession

of pertinent knowledge. At this point, the structure of information is examined. The levels of analysis are presented as follows:

REALITY ↔ elements of inform. level 0 ↔ categories of cognition ↓

[down] words/express./figures level 1 ↔ code of vocabulary letters/numbers/graphs level 2 ↔ code of signs

conf./sequen. of signals level 3 ↔ code of signals

These three (1, 2, 3) levels represent the shape of information, i.e., something material. The level dealing with lingual description, parametrical expression, or graphical presentation is accepted as the next level and labeled (1). The level of analysis that consists of letters in written form or syllables in a spoken language, as well as numbers, digital symbols, and visual representation, such as, graphs, ideograms, etc., are accepted as the next level and labeled (2). The last distinguished level of analysis is called level of signals and labeled (3). Because the shape of information is transmitted by signals in the telecommunication channel, the most popular concept of information pertains to a signal[234] as a unit of information. This leads to the conclusion that more signals equal more information.

Regarding the issue of the quantity of information, our thesis claims that this can be clarified with respect to the level of analysis. Analysis of a certain structure of information on the 0 (zero) level enables distinguishing of a certain number of elements of information within the scope of horizontal projection.

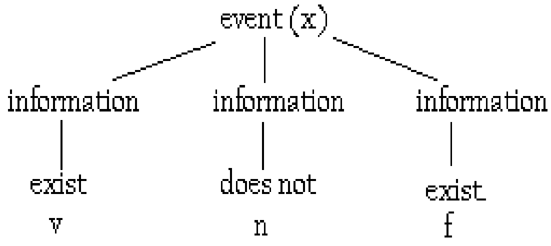
The fact that many different codes exist in relation to the three (1, 2, 3) distinguished levels gives rise to situations in which a given number of words/ expressions/ figures or a quantity of letters/numbers/graphs or a number of sequence of signals will be different, although the number of elements of information will remain the same.

Basic concepts

Existence of information is superior, with regard to either truthfulness or falsity; thus, nonexistence is the negation of either true or false information.

---

234 SIGNAL, electric impulse, light impulse, electromagnetic wave. Attention! Because a signal is a specific physical phenomenon that: 1) is generated by a certain system or transmitter for the purpose of conveying information; 2) is generated spontaneously by physical objects/processes from the surrounding world. It should be mentioned that the object of our interest in this work is the signal that is generated intentionally.



It is assumed that truthfulness (v - verum), nonexistence (n - nullus), and falsity (f - falsum) are states that make up a universal base (foundation) for analyzing information from the valued point of view. By analogy, in physics one has voltage (+), lack of voltage (o), or negative voltage (-). Similarly, in mathematics there are positive numbers (+), zero (o), and negative numbers (-). The above analysis provides a basis for replacing a proposition (sentence) with a segment of information in logical matrices, and likewise extending these for three values instead of two.

Truth-Functional Connective

The truth-functional connective for the three values (v, n, f)[235] - could be presented as follows. If there is one argument (p), then the matrix of negation will contain the following possibilities:

p	np^1	np^2
v	f	n
n	v	f
f	n	v

If p is v, then its negation is f and n; if p is n, then its negation is v and f; if p is f, then its negation is n and v.

If there are two arguments (p, q), then the combined matrix for alternative "v", [236]disjunction "/", conjunction "^", equivalence "≡", and implication "→" could be presented as follows:

p	q	pvq	p/q	p^q	p≡q	p→q
v	v	v	f	v	v	v

235 For a presentation of other concepts of multi-valued logic, see: Haack, Susan Philosophy of Logics, Cambridge University Press, 1978, ch. 11, pp. 205-220; also Czezowski, Tadeusz, Logika, Podrecznik dla studiujacych nauki filozoficzne, PZWS, Warszawa 1949, ch. 3, pp. 193-204. There, the author presents the concept and matrix of Lukasiewicz's 3- valued logic.

236 The functions of alternative and disjunction are presented here separately, see: Johnson, W. E., Logic, Dover Publications, Inc. 1964, part 1, ch. 3, p.32-33, and also Czezowski, Tadeusz, Glowne Zasady Nauk Filozoficznych, ZNIO - W, Wroclaw 1959, ch. I, pp. 68-69.

v	f	v	v	f	f	f
v	n	v	f	n	n	v
n	v	v	f	n	n	v
n	n	n	n	n	v	v
n	f	f	v	n	n	v
f	n	f	v	n	n	v
f	v	v	v	f	f	v
f	f	f	v	f	v	v

During the realization of a certain task, if **choice, elimination, coordination, estimation,** and **inference** were applied, then the above dependencies indicate what the value of these functions would be for a certain combination of two arguments. From this premise, logic in the most general manner provides guidance concerning what would be expected as a result of thought, i.e., the dynamics of the process of thinking.

ALTERNATIVE

If there are two arguments (p, q), then the matrix for the alternative would be presented as follows:

7 -

	p	q	p v q	comments
1	v	v	v	suffices → for a choice of true argument
2	v	f	v	does not exclude → a choice of true argument
3	v	n	v	does not exclude → a choice of true argument
4	n	v	v	does not exclude → a choice of true argument
5	n	n	n	excludes → a choice
6	n	f	f	excludes → a choice of true argument
7	f	n	f	excludes → a choice of true argument
8	f	v	v	does not exclude → a choice of true argument
9	f	f	f	excludes → a choice of true argument

If **choice** is applied during the realization of a certain task, then the above dependencies show what the value of the alternative would be for a certain combination of two arguments.

DISJUNCTION

If there are two arguments (p, q), then the matrix for disjunction would be presented as follows:

	p	q	p / q	comments

1	v	v	f	Excludes → elimination of false argument
2	v	f	v	does not exclude → elimination of false argument
3	v	n	f	Excludes → elimination of false argument
4	n	v	f	Excludes → elimination of false argument
5	n	n	n	Excludes → elimination
6	n	f	v	does not exclude → elimination of false argument
7	f	n	v	does not exclude → elimination of false argument
8	f	v	v	does not exclude → elimination of false argument
9	f	f	v	suffices → for elimination of false argument

If **elimination** is needed during the realization of a certain task, then the above dependencies indicate what the value of disjunction would be for a certain combination of two arguments.

CONJUNCTION

If there are two arguments (p, q), then the matrix for conjunction would be presented as follows:

	p	q	p^q	comments
1	v	v	v	suffices → for true coordination
2	v	f	f	excludes → true coordination
3	v	n	n	excludes → coordination
4	n	v	n	excludes → coordination
5	n	n	n	excludes → coordination
6	n	f	n	excludes → coordination
7	f	n	n	excludes → coordination
8	f	v	f	excludes → true coordination
9	f	f	f	excludes → true coordination

If **coordination** must be applied during the realization of a certain task, then the above dependencies represent the value of conjunction for a certain combination of two arguments.

EQUIVALENCE

If there are two arguments (p, q), then the matrix for equivalence would be presented as follows:

	p	q	p ≡ q	c o m m e n t s
1	v	v	v	suffices → for true estimation
2	v	f	f	excludes → true estimation
3	v	n	n	excludes → estimation
4	n	v	n	excludes → estimation
5	n	n	v	suffices → for true estimation
6	n	f	n	excludes → estimation
7	f	n	n	excludes → estimation
8	f	v	f	excludes → true estimation
9	f	f	v	suffices → for true estimation

- 9 -

If **estimation** needs to be applied during the realization of a certain task, then the above dependencies illustrate what the value of equivalence would be for a certain combination of two arguments.

IMPLICATION

If there are two arguments (p, q), then the matrix for implication would be presented as follows:

p	q	p→q	c o m m e n t s
v	v	v	suffices → for true implication
v	f	f	excludes → true implication
v	n	v	does not exclude → true implication
n	v	v	does not exclude → true implication
n	n	v	suffices → for true implication
n	f	v	does not exclude → true implication

f	n	v	does not exclude → true implication
f	v	v	does not exclude → true implication
f	f	v	suffices → for true implication

If **inference** is applied during the realization of a certain task, then the above dependencies represent what the value of implication would be for a certain combination of two arguments.

Truth-Functional Connective (Comments)

When one analyzes the issue of **lack** and its meaning in an abstract manner, two different situations arise: first, when such information does not exist **at all**; and, second, when a certain subject **does not have** the necessary information at a particular time and place, but nevertheless has to make a decision and move forward. Truth-functional matrices apply to both situations. This involves a different approach than the one provided by traditional logic. Two-valued logic is based on each one actual or universal relation between a subject and the object of examination, whereas 3-valued logic refers to the information about this examination.

The simplicity and beauty of dichotomous division has enabled it to maintain a prominent position in logic since the time of Aristotle. As a guide to thought, however, such a division appears to be insufficient in situations in which steps are undertaken that are partially based upon a lack of information or even upon obvious ignorance. Examples that come to mind are the construction of airplanes or electronic computing machines. What could have been predicted over 60 years ago about computer applications? As never before in history, the actual virtual reality of information confronts a subject with a situation in which one has either true information, false information, or no information. Moreover, it is irrelevant if such information does not exist or if it is merely inaccessible.

Amount of Information

In the previous section, matrices of 3-valued logic were presented that depict an abstract situation in which, in addition to true and false information, the realization of a certain task takes into account a lack of information as a necessary component of logical relation. Now it would be constructive to return to one of the key questions related to the **quantity** or **amount** of information. Here we will refrain from discussing

Shannon's concept of **quantity** of information, which is based upon transferring signals by wire, because this concept is commonly known.

There is another aspect of the value of information that should be considered. Whenever an event x occurs, it is named an element of 'reality' (er) that could be presented by one element of information (ei), and then the following possibilities occur: First (p^1), that the representation is true (v); second (p^2), that the representation has not been achieved (n); and, third (p^3), that the representation is false (f).

	p^1	p^2	p^3
er	v	n	f

How many possibilities can be found if the object of description contains more than one element? If there are two (er^1, er^2) elements of 'reality', then the number of possibilities (p) is nine.

	p1	p2	p3	p4	p5	p6	p7	p8	p9
er^1	v	v	v	n	n	n	f	f	f
er^2	v	f	n	v	n	f	n	v	f

This means that the result of one representation must be equal to one of nine possible results, which, however, does not mean that the one completely truthful result is necessarily found among the nine representations. The reader should recognize this table as a **horizontal layout of the matrix of conjunction.** If there are three (er^1, er^2, er^3) elements of 'reality', then the number of possibilities is twenty-seven; if there are four elements, then the possibilities are eighty-one. Because of the possible increases in relation to the number of elements of 'reality', as described above, which has a power nature, the following formula is utilized instead of matrices:

Number of possible results of a process (in this case cognition) for 'reality' with x elements $N = 3^x$

Process of Cognition

The assumption is being made that the cognition of 'reality' is realized by means of **identification** or **measurement**[237]. Identification refers to quality, whereas measurement has to do with the characteristic of quantity. The realization of identification and measurement discloses the existing relationship between the quality and quantity of a certain object. The relationship between quality and quantity represents the structure of information. This can be described as follows:

measurement	n	(numerator)	quantity	ei^1
relation	-	(connector)		ei^2

237 The most general categories utilized in the field of philosophy are substance, quality, and relation. See; Harre, Rom , The Philosophies of Science, ch. 4, pp. 100-101, Oxford University Press, London 1972. However, for a more detailed presentation of this concept better serving our purpose, the categories of quantity, quality, and relation must be considered.

identification d (denominator) quality ei^3

The value of information is a function of method (Me)[238], technique (Te), and goal (Go). If, within the scope of the tasks of identification or measurement, method (Me), technique (Te), and goal (Go) are in agreement regarding the object (Ob) of examination, adequate information (equal within assumed scope) is obtained. Such information is then accepted as true information:

if er = ei then ei is true (v)

The principal issues within the scope of tasks of identification or measurement are a lack of, or an inconsistency of, method (Me), technique (Te), and goal (Go) in relation to the object (Ob) of examination. What would then transpire is either a lack of or an improper result. In other words, the result would be a lack of information or inadequate information. Inadequate representation would be considered false information:

if er = 0 then is null (n)
if er ≠ ei then if false (f)

When considering these components (Me, Te, Go) during the process of cognition, the following values are applied: (+), meaning proper; (o), meaning lack; (-), meaning improper.

These symbols are deemed suitable for describing the situation when the said components (Me, Te, Go) must act together. It seems more desirable to use the word **proper** than the word **true** when one is engaged in applying a method (Me), using a technique (Te), or pursuing a goal (Go). Thus, a slightly modified **matrix of conjunction** is utilized for the three arguments. The next table (page 14) shows possible variants of cognition focused on an object (Ob) as related to method (Me), technique (Te), and goal (Go).

Possibilities 1-27 can be classified as follows: The first possibility represents
a proper result. This can be interpreted as the **true** cognition of identification or measurement. A lack of result is represented by 19 possibilities, namely, 3, 5-8, 10-19, 21-22, and 25-26. An improper result is represented by 7 possibilities, i.e., 2, 4, 9, 20, 23-24, and 27. Even the lack of one factor causes a lack of result for the entire cognition. An improper result occurs when certain combinations of (+) and (-) arise for all factors.

---

238 The actual level of development in a given field of knowledge, which is based upon a confirmed theory, is represented by a method. For this reason, the value of information is changing over time, which, in turn, means that truth depends not only upon correct logical procedure, but also upon progress in a given field of knowledge.

	Me	Te	Go	Result
1	+	+	+	+
2	+	+	-	-
3	+	+	o	o
4	+	-	+	-
5	+	o	+	o
6	+	o	o	o
7	+	-	o	o
8	+	o	-	o
9	+	-	-	-
10	o	+	+	o
11	o	+	-	o
12	o	+	o	o
13	o	o	+	o
14	o	o	o	o
15	o	o	-	o
16	o	-	o	o
17	o	-	+	o
18	o	-	-	o
19	-	o	+	o
20	-	+	+	-
21	-	+	o	o
22	-	o	o	o
23	-	-	+	-
24	-	+	-	-
25	-	o	-	o
26	-	-	o	o
27	-	-	-	-

The following three possibilities, indicate that the second (++-) method (Me) and technique (Te) were in order, but the goal (Go) was inconsistent **(teleological error)**; the fourth (+-+) method (Me) and goal (Go) were in order, but the technique (Te) was inconsistent (**technical error**); the twentieth (-++) method (Me) was inconsistent, whereas the technique

(Te) and goal (Go) were in order (**methodical error**). In the aforementioned three instances, cognition resulted in **false** information because one of the three necessary components was improper.
With regard to both identification and measurement, the obtained truth has relative character. In addition to the function of method (Me), technique (Te), and goal (Go), other components include time (Ti) and place (Pl). During the time of Ptolemy, the widely held view of the Solar Planetary System was different from the one following the time of Copernicus. The same can be said of the measurement of the distance between the moon and the earth: What was believed a hundred years ago is different than what is believed today.

Conclusion Regarding Fundamentals

The previous three sections of this paper disclosed three different aspects of understanding the concept of true (v), a lack of (n), and false (f) information. The first was the logical aspect, depicted in the form of matrices, which indicated in a most abstract manner what the result of our thinking would be for two arguments
(p, q) if we considered three values (v, n, and f). The second was the probabilistic aspect, presented in the form of tables and formula, which illustrated what could have occurred during the time when information originated, i.e., what the result could have be (v, n, f). This section also considered a different concept of quantity or amount of information by modeling how a number of cognizable elements translated into a number of possible results.

The third was the corresponding aspect**.** The name here being borrowed from the correspondence theory of truth[239]. This aspect consists of three categories—quality, quantity, relation—and, in a general way, depicts with participation of what components: method (Me), technique (Te), goal (Go), the value of information (v, n, f) originate.

Before proceeding with the presentation of the structure of information and the spaciousness of thought, the author requests that the reader stop for a moment and try to imagine these three aspects of comprehending value (v, n, f) as a certain three-dimensional, geometrical figure assembled from three equilateral rectangles, of different colored sides, inserted into a transparent sphere, connected by a common point and adjusted for 120 degrees. Such an illustration is proposed because it appears to represent the nature of analyzed values.

---

239 See: Angeles, Peter, A., Dictionary of Philosophy, Barnes & Noble Books, New York 1981, p. 298. Also The Correspondence Theory of Truth, Stanford Encyclopedia of Philosophy (Internet)

Structure of Information (Horizontal Projection)

The structure of information, in horizontal projection**,** on the zero (0) level of analysis can be presented in the following manner. There is a relationship between an element of information and 'reality', on the one hand, and an element of information and corresponding categories of information, on the other hand. This is a general matrix consisting of object (Ob), method (Me), technique (Te), goal (Go), subject (Su), time (Ti), and place (Pl). These categories would need to be utilized for any type of examination. This also means that they become the **vectors of the virtual space of information**. With the exception of covering the totality of representation, the structure of this space assumes an efficiency for the servicing of information. Here, individual thinking encounters its social foundation, allowing an objectivization of its result. The category of subject (Su), which personifies the usage of method (Me), technique (Te), and goal (Go), is necessary for processing. An equivalent to the subject could be a computer functioning as an autonomous causative system. The last two categories refer to and describe time (Ti) and place (Pl**),** when and where**,** particular information originated.

Ob(v1)	Me(v2)	Te(v3)	Go(v4)	Su(v5)	Ti(v6)	Pl(v
object	method	technique	goal	subject	time	pla
what	how	what with	why	who	when	wh
ei^1	ei^2	ei^3	ei^4	ei^5	ei^6	ei^7

According to our thesis, these categories are necessary components of thinking. And regardless which process of thought is being examined, it will always be constructed in a dynamic mode based upon these components.

Spaciousness of Thought

In the opening section of this paper, the **vertical structure** of information was described along with the concept of **levels of analysis**. This concept served to examine what is manifested in the shape of information. Now, by delving deeper into the virtual realm of thinking, we next focus upon the well-known concept of **hierarchical levels**. Thus, when adjacent levels are involved, the level above represents the

location of something general (term, information), whereas the level below stands for the location of something particular (term, information). In addition to the relations between two levels during the thinking process, we also encounter the relation of something (term, information) with something else on the same level, between two sides that are named **coordinal** sides. Because something could represent inference, we therefore introduce the additional concept of consequential states, which applies to **diagonal** relations.

The logical frame of thinking appears in a spacious manner on hierarchical levels:

between two hierarchical levels,	- vertical	↑
between two coordinal sides,	- horizontal	→
between two consequential states,	- diagonal	.

The above instances do not pretend to fully describe the logical frame of thinking; they merely aim to demonstrate its spaciousness. In conclusion, we can say that spaciousness of thought manifests itself virtually in the following dimensions: vertically, between the levels; horizontally, between the sides; and, diagonally, between the states.

Levels of Synthesis

In order to make a general analysis of the contents of thought in vertical projection, it is necessary to introduce another category of levels that we call **levels of synthesis**.

These levels ascend from the zero (0) level of elements of information to parts, numbers, and relations at the first (1) level. And, from here to configuration and sequence, set on the second (2) level. Finally, from the second level (2), one moves to the structure of the model on the third (3) level.

	structure of model	level	3	↔ HYPOTHESIS
	config./sequence/set	level	2	
	parts/numbers/relations	level	1	
‘REALITY’ ↔	elements of information	level	0	↑ [up]

Please note that the numbers of the **levels of synthesis** are delineated by underlining.

Planes of Realization

Leaving aside the presentation of elements of information in vertical and horizontal projection[240], we next focus on the planes of realization. The thought that goes through the mind—whatever it may be—has a particular meaning and is therefore connected to the proper word. It also has specific parameters that allow connections among the various elements of information. Similarly, the mind must take the value of the involved factors into consideration. Nevertheless, if any measurement is involved, computation is assumed.

The **lingual** plane contains term, relation, and definition, and it serves to communicate results.

The **informational** plane contains object, method, and goal, and it serves to achieve transformation.

The **logical** plane contains alternative/disjunction, conjunction /equivalence, and implication/negation, and it serves to supervise evaluation.

The **functional** plane contains addition/subtraction, multiplication/division, and equation/inequation, and it serves to control computation.

The above relationships can be presented in the form of a table:

Lingual	Informational	Logical	Functional
term	object (Ob)	alternative/disjunction	+/-
relation	method (Me)	conjunction/equivalence	x/÷
definition	goal (Go)	implication/negation	=/≠

Conclusion

The dominance of computer science over the interpretation of many crucial problems related to the working of the **brain**, in addition to discoveries of important physiological characteristics of gray matter, has, in the author's opinion, created a perspective for understanding the thought process that is both limited and too general. The author believes that the philosophical sciences must provide a healthy counterbalance by focusing attention on specific rational activities of the **mind**, i.e., on thinking and reasoning. Alfred Whitehead wrote, "Philosophy is not one among the sciences with its own little scheme of abstractions which it

240 More detailed information a reader will find in my book: The Geometry of Thinking. An Exploration of the Multidimensional Space of Information.

works away at perfecting and improving. It is the survey of sciences, with the special objects of their harmony, and of their completion.[241]"

The limited scope of computer science may be attributed to its perception of objects of investigation via the category of quantity, thus implying that any resolution should be available via computation. At this point, it is instructive to quote Ray Kurzweil, who is a prominent figure in the field of artificial intelligence: "Our human intelligence is based on the computational process that we are learning to understand. We will ultimately multiply our intellectual powers by applying and extending the method of human intelligence using the vastly greater capacity of nonbiological computation."[242] It is obvious that, to a certain extent, computers have the ability, via their computational process, to imitate or replace the process of thinking. However, it must be maintained that the proper description of the nature of these processes requires a more diversified approach.

241 Whitehead, Alfred North, Science and the Modern World, The Press, New York 1953, Chapter V, p. 87

242 Kurzweil, Ray, The Singularity Is Near, Viking, New York, 2005, pp. 652, Chapter: The Limits of Computation, p. 128.

FACING THE COMPUTER.

SOME TECHNIQUES TO UNDERSTAND TECHNIQUE

LIESBETH DE MOL

Department for Philosophy

Centre for Logic and Philosophy of Science

tel: (00)32-(0)9-264 39 79 elizabeth.demol@ugent.be

In a very short time the computer developed from a mathematical-logical object into an everyday physical object upon which whole societies rely. Possibly due to its significance in our society, people have a rather ambiguous attitude towards computers: on the one hand, they consider the computer as something which should not make mistakes, it should be perfect. On the other hand, while considered perfect, they understand it as a stupid machine – the only thing it does and must do is calculate. This ambiguity is not only apparent on the level of everyday life, but is maybe even more explicit on the level of scientific research. Since Church and Turing stated their respective theses, hundreds of research papers have been published in which attempts are made to beat the *physical* Church-Turing thesis (thesis P), and probably as many papers have been published trying to defend it. Inspired by Martin Heidegger's *Question concerning technique*, the main purpose of this paper is to propose some strategies which might lead one to a better understanding of what a computation actually is. Starting from this position, the discussion on the yes/no possibility of physically beating a Turing machine will be questioned.

Introduction

In a very short time the computer developed from a mathematical-logical object into an everyday physical object upon which whole societies rely.

Possibly due to its significance in our society, people have a rather ambiguous attitude towards computers: on the one hand, they often get angry with their computer because it crashes or because it does something they did not ask for. They consider the computer as something which should not make mistakes, it should be perfect. On the other hand, while considered perfect, they understand it as a stupid machine – the only thing it does and must do is calculate. This ambiguity is not only apparent on the level of everyday life, but is maybe even more explicit on the level of scientific research. Since Church and Turing stated their respective theses, hundreds of research papers have been published in which attempts are made to beat the *physical* Church-Turing thesis (thesis P), and probably as many papers have been published trying to defend it.[243] This physical version of the thesis states that not only the computer, but every effectively realizable physical system can be defined in terms of Turing machines.

Inspired by Martin Heidegger's *Question concerning technique*[244], the main purpose of this paper is to propose some strategies which might lead one to a better understanding of what a computation actually is. Starting from this position, the discussion on the yes/no possibility of physically beating a Turing machine will be questioned. While this discussion is of course a very important one – both from a scientific as well as from a philosophical point of view – it will be argued here that it might also be interesting to shift attention from the question of what might (not) be computable to the question of what a computation actually is.

A question concerning Technique

In 1936 Church and Turing each independently proposed their respective theses. If these theses are true then it naturally follows that those decision problems not solvable by any Turing machine, or any other equivalent formulation, are non-computable problems. Since then hundreds of other decision problems have been shown to be unsolvable – problems which have not remained restricted to the domain of mathematical logic, even in theoretical physics there exist several unsolvable decision problems.[245] Despite or due to these results many authors claim that it might be possible to beat thesis P: they are convinced that there are physically realizable processes able to solve

243 The physical Church-Turing thesis however, should be neatly distinguished from its original mathematical formulation.

244 The official English translation of the title of this essay is: Question concerning technology. This translation is considered as a bad translation here, since Heidegger himself explicitly states in this essay that technique should not be considered as something technological.

245 Some authors mistakenly use such results in arguing against thesis P. See De Mol 2006b.

these decision problems.[246][247] Others argue that this will never be possible: the limit proven to exist by Turing is not only a theoretical but also a physically existing limit .

While a solution to the truth or falsity of thesis P would of course be interesting in itself, one must admit that the approaches both of pro and contra have not been able so far to provide for any definite answer. Motivated by this fact, the question must be raised whether it could not be more interesting to leave open this problem for a while, and focus on what this problem is actually about. When originally posed in 1936, the Church-Turing thesis had no link whatsoever to the question of how to outrun a computational process. It was about the question: what do we mean *exactly* when we say that something is computable? This question is not only significant in relation to the ongoing discussion on thesis P, but is maybe even more significant from a philosophical point of view, especially if the materialization of the subject of this question is taken into account: the computer.

The computer has become an omnipresent object in our society: the variety of applications and our dependencies on them could have hardly been imagined by Turing when he first described the abstract complement – the universal Turing machine – of this general-purpose machine. From its first use on – making calculations for the A-bomb – it was clear that these applications are not merely restricted to the field of scientific research, resulting in a "technology" that affects every man in the street.

In this respect Martin Heidegger's question concerning technique becomes an insistent one. Although this poetical text is often interpreted as a critique on the "technification" of modern society it can also be understood as an appeal to man to become aware of the "essence" of the "technique" present in society – technique being the way man perceives, acts in and with the world.

Since calculation and control of information are two typical "features" of the way man perceives, acts in and with the world nowadays, the computer can be understood as *a* physical realization of this "technique". Relating the question concerning technique, as posed by Heidegger, to computers implies that it is fundamental that man does not simply uses

---

246 In Cotogno 2003 one finds an overview of many different proposals for hypercomputation (including physical supertasks and infinite computations; interactive systems; analog computations and quantum computations), together with arguments which show that none of these is able to "beat" thesis P.

247 Without wanting to argue against the possibility of "beating" thesis P, the author would like to add one important question: Suppose that one would have constructed a machine M for which it is claimed that it outputs a non-recursive function, how will one verify this empirically? How will we humans be able to observe this? This is impossible, since the "thesis that M computes a recursive function is consistent with any finite chunk of data." (See Shipman, 2000.)

the computer as a mere instrument, but that he knows what he is using and how he uses it:

> Darum liegt alles daran, dass wir den Aufgang bedenken und andenkend hüten. Wie geschieht dies? Vor allem anderen so, dass wir das Wesende in der Technik erblicken, statt nur auf das Technische zu starren. Solange wir die Technik als Instrument vorstellen, bleiben wir im Willen hängen sie zu meistern. Wir treiben am Wesen der Technik forbei. (32)

Indeed, as long as we do not see that the computer is not merely an instrument but rather a realization of what technique is, its essence remains hidden and can thus form a danger:

> Das Gefährliche ist nicht die Technik. Es gibt keine Dämonie der Technik, wohl dagegen das Geheimnis ihres Wesens. Das Wesen der Technik ist [...] die Gefahr. (27-28)

The fact that the majority of mankind uses many of the possibilities of the computer, without ever knowing what is beyond the monopolized interface, is just one concrete example of this problem. Indeed, the only reason for this monopoly being possible is the fact that most of the people never get beyond their GUI.[248] The only way out is to face oneself with technique as it is.

Strategy I : Post's Machine

In a little booklet called "Post's Machine" the Russian mathematician Uspensky describes a "toy machine" – first described by Emil Post in 1936 – to show how one can advance abstract concepts such as "algorithm", "universal computing machine" and "programming" for school children.[249] After having explained the inner workings of a Post machine, Uspensky gives a long exposition on how to "program" the basic operation of "+ 1" on a Post machine and shows in this way that this for us seemingly trivial operation becomes far from trivial when implemented on this machine. Indeed, instead of one step, the machine needs 23 instructions to perform this simple calculation. Furthermore, making it run one might need hundreds of steps to add 1 to a number, depending the initial condition. This example clearly shows how that which is considered as a basic step in arithmetic becomes in itself a complex operation when performed on a medium which was constructed with the purpose of formalising the intuitive concept of a computation. In other words, in working with this machine something very fundamental about computations – and thus technique – is understood: when we perform calculations in our everyday life, this is merely *a* way to do it. That, which seems to be the most basic operation in a rather absolute way, is not

---

248An existing strategy to counter this problem is proposed by the Free Software Foundation, which makes it possible not only to freely distribute the software itself but also the source code. In this way you can e.g. change the code when you want things to be done in another way – a freedom which is unthinkable for the average windows user.

249 See Uspensky 1983.

what it seems to be. It was exactly this kind of understanding that was fundamental to Church's and especially Turing's argumentation with respect to their theses: in order for their arguments to work, they had to analyze the possible processes underlying our everyday way of e.g. adding two numbers.

The idea of teaching schoolchildren that the way they normally calculate is not the only way, and certainly not the way their computer calculates, together with the fact that in this process they can also learn to understand some basic problems of programming, should at least be taken into consideration in the light of Heidegger's question concerning technique: it is one way to go beyond the GUI. Especially in the light of the omnipresence of the computer in our society and its consequent economical value, the idea that people don't have a clue of what a program is or how their computer works can at least be called disquieting.[250]

Strategy II: How Computers Constitute New Branches of Science.

As was stated earlier, technique is the way man acts in and with the world – his means of communication. This implies that technique is the framework through which we perceive the world in a certain way. Since technique is on the one hand reflected in the objects man creates, and is on the other hand the framework through which we perceive the world, the "technical" world must in its turn influence our thinking and understanding of the world, since we are part of it. Understanding that technique – when understood as the "technical" world – is not something passive, solely made for man's needs, simply standing there being available, but rather something that influences and changes man in a fundamental way, that there is a reciprocal relation between man and that which he makes, is considered as basic to our understanding of technique. But in what way does the computer and its calculations, as an explicit materialisation of "technique", influence our thinking and understanding?

There are many examples – the computer has touched upon almost every aspect of our life, ranging from war to art – but the ones focused on here are motivated by a lecture given by Von Neumann in 1949.[251] After having argued for the necessity of building up an intuition of a given mathematical problem one is investigating, he remarks that in some cases the computer might be the only way to build up such an intuition. Indeed, it might e.g. be possible that such intuition is blocked off because we simply cannot perform enough calculations in a reasonable time. It is in this respect that the computer has changed and even founded new disciplines of science. It has become possible to build up an intuition of

250 It should be noted though that no paradigm should be allowed to dominate education.
251 See Von Neumann 1966.

certain problems which would not have been possible before, and in some cases, the problem (and the intuition of it) even only became a problem, arising because of (the use of) the computer.
One of the most celebrated examples for which the development of the computer has been fundamental is fractal geometry, which is in its turn closely related with another such discipline, chaos theory. Neither of these branches of mathematics – with applications in other sciences – would have been possible without the help of the computer. Another example, intimately connected with the rise of the computer, is computational complexity theory, which asks questions about time and space complexity of decision problems and is fundamental with regards to the safety of the internet.
Nowadays, there are concurrent models for computations over the reals because of the explosive use of the computer in physics. Since physics works with the continuum and the classical model of computation is discrete it seemed necessary to develop such models. The fact that even new mathematical theories were developed for handling the problem of scientific computing, clearly shows that the computer now plays a vital role in physics. In order to study certain problems one no longer sets up a traditional experiment. Instead one makes a model or a simulation of the problem at hand, for which one can easily change any parameter in a couple of minutes or even seconds in order to study the behaviour of a certain physical process in a more general way. An interesting example is the research done on cellular automata (CA), where one of the big names nowadays is Stephen Wolfram.
CA's are mathematical objects for which it was shown that they can calculate anything a universal Turing machine can calculate. Although being abstract formalisms, just as λ-calculus, CA have not remained restricted to the domain of mathematical logic and theoretical computer science. They were developed by Von Neumann in collaboration with Ulam, as mathematical models of self-reproducing artificial systems. Nowadays, CA's are studied in the field of theoretical biology, and are one of the frameworks of artificial life. Without going into further details here, it is important to note that this research on CA often comes down to the "simulation" of (certain aspects of) physical systems. Stephen Wolfram for instance gives several examples in his highly debated book *A new kind of science* of CA-like models which simulate e.g. the growth of plants, pigmentation on certain organisms, and fluidic phenomena. Another example is Langton's ant, which exhibits a dynamical transition from "chaotic" to periodic behaviour. Both Langton and Wolfram are convinced that it is possible to simulate "life" in simple models such as CA and Turing machines and they thus have to assume the truth of thesis P.[252]

252 This is even very explicitly stated in Wolfram 2002, where it is presupposed that nature is algorithmic.

A Battle between Nature and Computing Machines?

In 1986 Langton published a paper on artifical life in which he stated:[253]

> The ultimate goal of the study of artificial life would be to create 'life' in some other medium, ideally a *virtual* medium where the essence of life has been abstracted from the details of its implementation in *any* particular hardware. We would like to build models that are so life-like that they cease to be *models* of life and become *examples* of life themselves." (147)

Indeed, to Langton (and the same goes with Wolfram) it is not correct to speak of simulations of life: it is "real" life. This is comparable to the strong AI position in which it is stated that it is possible for a computer not only to simulate intelligence but to be *really* intelligent. Both hard AI and alife clearly presuppose thesis P. As was already stated in sec. 2 however, the question must be raised whether it could not be more interesting to leave open the problem of the truth or falsity of thesis P and instead focus on what this thesis is actually about. In its original form – the Church-Turing thesis – the question was not how can we create life in a formal system, or how does life outrun a formal system, but rather: what is a computation?

Both defenders and opponents of thesis P never seem to go back to the original papers by Church and Turing from 1936 – in the best case they are reduced to the level of a non-read reference – and consequently never *really* reconsider this last question, although it is fundamental to their work. Indeed, there seems to be only one important direction taken into consideration in this ongoing discussion: how do we go (or can we never go) from nature to computations? Focus is always put on the physical processes themselves and how they can or cannot be implemented on a computer, the other side of thesis P is hardly investigated within this domain. Identifying physical processes with their simulation in CA or the opposite idea that there must be something "non-computable" about physical processes – whatever that may be – then become mere symptoms of this one-directedness.

But why should one bother about this imbalance in the discussion on thesis P? There are two reasons. First of all it is more than significant that people are aware of what technique is. Since calculation is one of the features of modern technique, trying to use it as a way to capture nature or to differentiate it from technique (as calculation), while not focussing on it, is a typical example of how technique gets hidden away: one uses its materialisation as a way to escape or control it. In this way, one will never be able to understand it.

The second reason is in fact an application of Heidegger's thoughts on truth. Without going into more details, it is important to note that this concept is based on a kind of Escher-like effect: in focussing on one

253 See Langton 1986.

thing, in putting something in the foreground, there is always a background, constitutive for the foreground. Although you cannot focus on two things at one and the same time, it is important to be aware of the fact that there is a background. Trivial as this principle might seem, it is not. Applying this idea to the discussions on thesis P, one has to ask the following question: In focussing on physical processes, ignoring the other side of thesis P, does one not risk to exclude some very interesting philosophical and scientific questions?

To give just one example, why does one seldom ask: are there things computers can do, which are not apparent in nature? There seems to be at least one very nice example here. As every programmer knows it is very important to have a good random number generator (RNG). Nowadays, every newly constructed RNG has to pass for several tests – performing only one is not good enough. It is in this context that Marsaglia developed the software: DIEHARD, a battery of tests of randomness. In the instructions for using DIEHARD, one reads:

> I hope you will inform me of results, good or bad, of new kinds of generators you have tested, particularly deterministic generators, but also the output of physical devices. (I have found none of the latter that get past DIEHARD, and would like to learn of any that do.) Since, in my opinion, there is no true randomness, collective experience in finding sequences that depart from the theoretical ideal in a significant way can perhaps lead to better ways for finding those that do not.

Indeed, while one would expect physical RNG's to be the best of possible RNG's, none of the ones Marsaglia tested got past his tests. On the other hand, there are several deterministic generators which do pass it. While this is of course not a valid proof of the non-existence of statistical randomness in physics, it is an interesting observation, which – as noted in the above quote – might progress certain research.

Strategy III: Playing with formalisms.

As was shown, instead of focussing on the physical side of thesis P, it might be interesting and even fundamental to further investigate the computations themselves. Even if one feels the urge to solve thesis P, the question poses itself in what way we can ever solve this problem if we have never looked at the behaviour of computations without superimposing anything concrete on them?

Of course computations are always captured in a form: CA, Turing machines, λ-calculus,....The ones focused on here are tag systems. They were developed by Emil Post in trying to find the most abstract form of symbolic logic – and abstract they are. As was argued in *De Mol 2006a*, Post's method in arriving at a variant of the Entscheidungsproblem in 1921 was to construct more and more general forms of mathematics. In this process of developing more and more general forms he ended up – at a given moment – with his form of tag. A tag system is defined as follows. Given an alphabet A of μ symbols, e.g. A = {0, 1} and a natural

number v, e.g. v = 3. With each of the letters of the alphabet there is a corresponding word over the alphabet. E.g.:

$1 \longrightarrow 1101$
$0 \longrightarrow 00$

Given an initial sequence, depending on its first symbol, tag the word corresponding with this symbol at its tail and then remove the first v elements. This process is repeated for every new string produced, until the empty string is produced. Here is an example, applying the above given rules:

100101010011
1010100111101
01001111011101
0111101110100
....

Now why should one be interested in tag systems? First of all it should be noted that they are "as complex" as CA, since there exist universal tag systems. Secondly, they seem to be good formalisms to allow for a further analysis of "computations" in the light of the above given discussion: it is hard to superimpose anything concrete on these systems since they were developed to avoid this! Moreover they are very easy to implement on a computer, and run very fast. Given the "meaninglessness" of tag systems it is possible to investigate computational systems in a less prejudiced way since there seems to be no "meaningful" interpretation at all: neither for their behaviour nor for their rules. So how should one start? The first thing to do is – to follow the words by Von Neumann – to build up an intuition of these systems. This can be done by setting up an experimental dialogue between tag systems, the physical machine they are run on, the programming language they are written in and the human wanting to understand them. You can then see what different kinds of behaviour tag systems can lead to in varying several parameters like v and µ. It is also important to work some systems out by hand, since performing the operations by yourself leads to other intuitions of the systems. But what other kind of research could one do? There are several interesting research questions here, but they cannot be discussed here.[254]

More significant here is the setting up of an experimental dialogue. If one ever wants to understand part of what technique is, one should not consider formal systems and their physical realization as something which is not part of the world – mere products of the human mind, standing at our disposal when we like it. Rather one should not forget that *as* products of the human mind, they are part of this world, and can be as

254 To give just one example: is it possible to prove that 2 symbolic tag systems, with v > 2 are universal and/or whether there exist examples in this class with an unsolvable decision problem. This question is significant since one seems to be in need of methods, different from the usual ones.

physical as anything else. It is in this sense that setting up an experimental dialogue between e.g. tag systems and a human being is one of the possible strategies to understand technique. In translating my questions to my computer, and waiting for the answer from the tag systems, one can only learn that they cannot be controlled by a human mind – although they were invented by one. If I ask a question, I am never sure of what the answer will be. Often, the answer I get is more like another question, posed in language which is not mine – and one is forced into further questions.

Furthermore in performing computer experiments, one quickly learns that the machine running the code does and must make mistakes due to its physical nature. Overflows and calculation errors are "bugs" one does not expect if one never goes beyond the traditional GUI. Going back to the beginning of this text, many people are indeed rather ambiguous towards computers, scientists included. This ambiguous attitude can only be resolved in understanding that computers can and must make mistakes and be limited. After all, a computer is a physical system obeying physical laws. Doing computer experiments, can show one that a computer does not "deserve" the special status of being perfect and stupid at one and the same time: it shows us that computers are much more "worldly" than is often believed.

Conclusion

What is technique? Is it something technical? No. Rather it is the way man perceives of, is in and with the world. Not being aware of the way you are in the world can at least be called problematic. To Heidegger, this is one of the biggest problems of modern society. We do not see in what way we are approaching the world we are part of. One way to make a start at an understanding of technique is to face oneself with one of the "features" of technique: computations which are physically realized in our world through the computer. In this paper, some strategies were proposed to allow for a better understanding of computers and computations. These were linked to the ambiguous attitude of man towards computers, more specifically linked to an existing ambiguous attitude of scientists.

But why does one need this obscurantist philosopher Heidegger here? It could have been perfectly possible to argue for the significance of a deeper awareness of what a computer/computation is without him. Indeed the mere omnipresence of the computer and our dependence on it would have been enough.

Heidegger's text is an appeal to man, written in a style that, when taken seriously, forces one to think. From this text one can begin to understand that it is far from trivial, and maybe even necessary, that one starts to

have a closer look at the “technical” world. In this way, Heidegger’s text, if read in a certain way, is yet another strategy to understand technique.[255] The main reason however for including Heidegger here, is the mere fact that most of the philosophers don’t have a clue of what a computer is. In the meantime they are discussing topics such as: the yes/no possibility of an intelligent computer, going beyond the Turing machine (or not), the notion of randomness and coincidence far away from RNG’s,... They do this without ever having taken the trouble of going through the details of Turing’s paper, let alone, following the motions of a universal machine, understanding what it is, and what it is not. Heidegger’s text was included here in order to make an appeal to the philosophers *through* the words of a famous philosopher, obscurantist though he may be. As Friedrich Kittler once formulated it:

> In this way, computers are sold whose architecture is not so much defined by the state of the art but by a pre-history or firm bureaucracy that crystallises into hardware right away. And if the ideal of software [...] would ever triumph, the bureaucratisation would be perfect: The hardware, in spite of its programmability, would irrevocably be obscured under its packaging. To stop this coincidence from happening seems to be an eminent political goal. If computers are the first machines to reduce the contingency or incomputability of some, though not all futures to a finite degree, its own contingency should remain as open as possible. [...] If somebody went and wrote all the programmes hitherto running under the name of philosophy into hardware, that would be the goal itself. (*Kittler 1987,* 131)

Bibliography

Paolo Cotogno (2003), *Hypercomputation and the Physical Church-Turing Thesis*, in: British Journal for the Philosophy of Science, **54**, pp. 181—223.

Liesbeth De Mol (2006a), *Closing the Circle: An analysis of Emil Post’s early work*, The Bulletin of Symbolic Logic, **12**, 2, 267-288.

Liesbeth De Mol (2006b), *Refocussing Unidecidability. Questioning some extensions of the notion of formal undecidability to other domains.* In: D. Aerts, B. D'Hooghe and N. Note (Eds.), Worldviews, Science and Us: Bridging Knowledge and Its Implications for our Perspectives of the World. World Scientific , Singapore, To appear in 2006.

Martin Heidegger (1962), *Die Frage nach der Technik* in: *Die Technik und die Kehre*, Neske, Tübingen. Originally published in: *Vorträge und Aufsätze*, 1954.

Christopher Langton, *Studying Artificial Life with Cellular Automata,* Physica D, **22**, 1-3, 120-149.

Friedrich Kittler (1987), *Hardware, das unbekannte Wesen*, in: Sybille Krämer (ed.), Wirklichkeitsvorstellungen und Neue Medien, 119—132.

255 It was in fact this text that led the author from philosophy to computers and programming.

Joe Shipman, *Physical Computability,* FOM list, 23 August, 2000. Available at: http://www.cs.nyu.edu/pipermail/fom/2000-August/004244.html.

Vladimir A. Uspensky (1983), *Post's machine*, Mir Publishers (Little Mathematics Liberary)", Moscow.

John Von Neumann (1966), *The General and Logical Theory of Automata*, Arthur W. Burks (ed.), University of Illinois Press, Urbana, London.

Stephen Wolfram (2002), *A new kind of science*, Champaign, Wolfram Inc.

TWO MODELS OF COMPUTING MACHINE AND THEIR EPISTEMOLOGICAL CONSEQUENCES

TERESA NUMERICO
Dept. Communication sciences

University of Salerno

Tel. +39 089 96 2220 or +39 340 77 31 242

t.numerico@mclink.it

In this paper I will illustrate two epistemological perspectives that were active and influential in creating the computer and defining its tasks. The first model was based on logic and its nature was more explicitly theoretical; the other model was based on a simulative analogical method, oriented towards the practical results, but it offered an implicit different epistemological perspective on machines, particularly in the perspective of the interaction between the device and the human beings. I will isolate the two influences in the light of some historical relevant issues of information technologies, with the aim of furnishing some hints for the future developments of the discipline.

The paper will show that the progress of the vision of computer as a communication device was not included in the logical perspective. Computer science is considered the product of the logic theoretical results of the '30s of the '900 and of the technological development due to the Second World War, however if we approach information technologies in order to establish their epistemological status and we want to describe their methods and scientific ancestors we find ourselves in a difficult position. What is the deep nature of computer science? Where does it take its theoretical and practical representational models of itself and of its object?

The logic model of the calculating machine

The first and most influential model of computing could be found in the tradition of the logic researches of the first 30 years of the XX century. It was based on the idea of creating a formal system, adequate to the representation of all mathematical knowledge. Alan Turing's negative result about the decidability of a formal system established a possible definition of effective computability and, at the same time, its failure when he solved negatively the decision problem in 1936. In establishing his negative result, however Turing introduced a *costruens pars*: the Universal Machine was a model of mechanical computability and could be used as a theoretical environment for the built of a practical machine, in spite of the fact that it defined the theoretical boundary to the possibility of solving problems by a calculating device (Turing 1937). When, in 1945, von Neumann created the first official project of the stored-program machine, known under the name of *First Draft* (von Neumann 1945), Turing's machine was the most effective available model and it was the inspiring idea of the logical structure of the practical computing machine.

Turing himself participated to another project for the creation of an electronic calculating machine in the United Kingdom, starting from 1945 (Turing 1945), and in describing the new machine in front of the audience of the London Mathematical Society, he explicitly stated the connection between the theoretical machine that he invented and the practical device[256] that, at that time, he was planning to build:

> Some years ago I was researching on what might now be described as an investigation of the theoretical possibilities and limitations of digital computing machines. I considered a type of machine which had a central mechanism, and an infinite memory which was contained on an infinite tape. This type of machine appeared to be sufficiently general. One of my conclusions was that the idea of a 'rule of thumb' process and a 'machine process' were synonymous. The expression 'machine process' of course means one which could be carried out by the type of machine I was considering (Turing 1947 in Copeland 2004: 378-379).

While establishing the connection between logic and his practical model of the machine, Turing clearly acknowledged that his theoretical machine represented both the possibilities and the limitations of the first practical devices that were built at that time. In this perspective logic was not only the solution of the computing machine problem, but was part of its

256 From the end of the Second World War Turing was involved in the British project of the National Physical Laboratory (NPL) for the building of an electronic machine called ACE (Automatic Computing Engine). His project for this machine was approved in 1946, but he completed the technical report needed for the approval at the end of 1945 (Turing 1945), see (Numerico 2005) for more details.

problems, inheriting all the limits established in the disciplines by Kurt Gödel and Turing himself, more than ten years before[257].

Logic and AI

The birth of the computer was connected with the project of creating a machine capable of intelligent behaviors. There were various tendencies to create such a machine. According to a survey view proposed by Claude Shannon in a lecture given in 1962, at that time there were three major solutions:

> One may divide the approaches to this problem [the problem of simulating the human or animal brain, at least from the behavioristic point view] into three main categories, which might be termed the logical approach, the psychological approach, and the neurological approach (Shannon 1963: 844).

However the so-called "neurological" approach was not enough developed at that time and it was also penalized in the following years by a drastic diminishing of research funds, the logic approach and the psychological one shared the common view that in order to achieve the accomplishment of an intelligent task by a machine it was necessary to give it some deduction capabilities, together with some heuristic decision strategies.

The project of Artificial Intelligence (AI), as it was expressed in the proposal for the official field starting conference in 1956[258], consisted in the belief that the lack of memory space was only a temporary obstacle for the success of the simulation of the 'higher functions of the human brain' and the problem of artificial intelligence was represented by the programmers' 'inability to write programs taking full advantage' of the machine. The scope of the field and its agenda was strictly linked to the previous logic studies, also because many of the major scientists in this area shared a high level training in logic (Aspray 1985). The project of the AI was mainly concentrated on the simulation, and consequently the substitution, of human functions by the machines, based on the assumption that some of the higher functions of the brain or of the mind, had no relation at all with their embodiment inside the human beings. Most of the work of AI relied on the functionalist attitude with regards to

257 There is no space in this paper for an extensive detailed description of the influence of logic over the computer, for more information on the epistemological and practical role of the logic and rationalistic approach in the birth and development of the computer see for example (Davis 2000).

258 I am citing here the Darthmouth Conference organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon, retrieved from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html on 17/09/2006

intelligence and its mechanical simulation. Moreover the early activities of AI researchers was mainly related to a very narrow conception of intelligence, based on toy-problem solving in restricted domains (Franchi and Güzeldere 2005: 46). As McCarthy clearly stated:

> A program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows (1959).

Common sense is identified with deductive capabilities that allow the computer program to put together the information it is given with the rest of the information it possesses in order to obtain all the relevant consequences of the facts it knows. This attitude clearly privileges the deductive method that is typically related to the classical conception of logic applied to common sense reasoning. From this perspective it was possible to simulate the higher functions of the mind by a calculating machine and this could be done by a sufficient injection of logic and deducting capabilities inside programs. Gps, Advice Taker, Bacon, Mycin and other well-known success programs were some of the positive case histories that fulfil this theory. We can summarize very briefly some of the most relevant ideas behind those AI projects:

- Intelligent programs are based on a lot of knowledge.
- Knowledge, written in advance, must be represented in programs.
- Mathematical logic is a good notation for writing the knowledge down (sometimes the only adequate notation).
- Axiomatic theories that represent knowledge should appear explicitly in the programs.

All these ideas rely on an interesting hypothesis that was never proved, or even explicitly stated, that most of human thought and of the consequent knowledge we acquire, is produced by deduction. This implicit premise was harshly criticized years later by one of the most important contributors to the AI research Drew McDermott (1987). Though the strong AI proposal had various difficulties after the '80s of last century, we have to admit that one of the most relevant sources of information technology and artificial intelligence must be identified in the formalistic logic tradition, including all its positive and negative results. The idea of the machine entailed by this approach was that of a calculating device, isolated from the environment and programmed to achieve all the relevant intelligent result, using the deductive method model in order to achieve the desired mechanical 'reasoning' results.
Even if this method was very common among many of the researchers who belonged to the AI field there were some relevant exceptions, such as Alan Turing. In the last years of his research on mechanical intelligence, he showed an increasing interest for a different and more complex approach to machines and declared that, in order to emulate intelligence with a mechanical device, the 'discipline' of drawing all

conclusions from the assumptions was not enough. It was necessary that the machine was capable of learning how to activate some “initiative” in its behaviour. The initiative could fail and end up with producing mistakes, but it could also obtain innovative and more interesting results (Turing 1948).
Although one of the most relevant AI objectives was the simulation of human intelligence, the major successes that researchers obtained in this field were in the creation of ‘intelligent’ interface between the machine and the users in which the device acted as “an ‘helper’ or ‘assistant’ or ‘prosthetic’ under the guidance and control of human beings rather than autonomous robot standing on its own wheels (or feet, technology allowing)” (Franchi and Güzeldere 2005: 46). It is along these lines that we can observe the development of the second approach to the concept of the machine.

The Memex of Vannevar Bush: the communication model of machine

If someone had told Vannevar Bush (1890-1974) that his most famous project in the twenty-first century would have been the Memex, he probably would have not believed it. He can be considered one of the most successful US scientists during the 1930s and the 1940s, not only for his astounding scientific and technological achievement, but also for his success as a politician and an administrator in science. Trained as an electrical engineer, he obtained the Ph.D. at MIT, where he became vice-president and dean of the school of Engineering in 1931. His success as bureaucrat, though, did not prevent him from obtaining a major results in science with the construction of an analogue calculating machine, the *Differential Analyzer*, created in 1936, it was the most powerful calculating machine during the War. In 1939 was appointed president of the Carnegie Institute of Washington and left his career at MIT. In 1940, at the beginning of the War, he presented to President Roosevelt a project for the creation of an organization for the development of critical technologies as well as cutting edge weapons with the help of scientists. The organization was approved and later called *Office for Scientific Research and Development* (OSRD), therefore Bush was at the center of a powerful network of scientists that accepted to cooperate with military partners during the war. He was also one of the creators of the peace time substitute of the OSRD, the National Science Foundation (NSF).
It was probably for his double experience as a scientist and as a technocrat that he developed the broad vision that brought him to address the library problem, in his Memex project. The famous article in which he launched the proposal was published at the very end of the War

in 1945, (the same year in which the *First Draft* by von Neumann started to circulate among the interested scientists) though he was clearly thinking about it even before. His analysis of research developments from a privileged position pushed him to envisage two major problems of the science making process of his times: specialization of scholar and the amount of literature produced in each area. It was almost impossible at his times to “keep abreast of current thought, even in restricted fields”. A record, in Bush’s opinion, in order to be useful in science, had to be continuously extended, stored and consulted. This approach could be assimilated to the idea of information as process that in order to be valuable needed to be used by people, that was shared by Wiener too. The difficulty in managing the scientific literature of every specialization has increased radically from 60 years ago. According to a survey made in 2003[259], it was estimated that the information stored on paper, film, magnetic and optical media has about doubled from 1999.

The library problem raised by Bush was not solved by the new digital technologies, his solution, though, was thought-provoking and is still inspiring for us now. He argued that selection was the key factor to deal with such a big amount of information and criticized the mechanisms commonly used by libraries to index information. It was not only a bare proposal for improving with mechanization the actual process used by libraries to organize bibliographic data and obtain outputs for specific researches, he was suggesting a complete change of paradigm in the access, retrieval and creation of information as well as in managing knowledge.

> Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome […]
>
> The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain (Bush 1945: 32-33).

The idea of simulating the associative strategy adopted by memory when it selects a trail of ideas in the mind was a new perspective in the information management field. Association was meaningful only for the mind that created it, but it could be very effective in retrieving information and making sense of raw data. Bush did not believe that machines could really emulate the human memory, but was convinced that machine could “augment” the natural power of the human brain in making sound and

259 Lyman, Peter and Hal R. Varian, "How Much Information", 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on 17/09/2006.

useful association. Memex was a device that, using analogue tools to store data, had the capability of creating trails of associations according to the selection of the user and to store them in its huge memory. Bush's machine had nothing to do with the huge stored-program digital calculating machines that were about to be built during the 1940s. It was a desk machine, designed for professional workers such as lawyers, physicians, chemists, historians that could store all the data useful for their work and retrieve them swiftly whenever they needed them. Though in 1945 Bush failed to recognize the development and potentialities of the new digital technology, never becoming an expert of it, his vision was in a certain sense cast over the digital technology to reach us and the need for "analogue" tools to deal with the information overload and the best interfaces used to access data in a digital world. In this sense he can be considered an inspiring precursor of the Web and the hypertextual world that was born with the creation of the personal computer and of the communication technologies.

Moreover Bush raised another key issue about what machines can do, in his paper *Memex II* that was at the center of many discussions during those years. His answer, differently from the first AI scientists, was that their main feature was the recollection of stored data, and that their logical equipment was only a secondary characteristics. Discussing about logic, he launched a strong attack against the abuse of it. In his view logic could be used only when the premises were precisely defined and data were clearly stated, without this guarantee logic was meaningless. The abuse of logic consisted in the application of Aristotelian rules to undefined premises. All sound conclusions that could be obtained from some correctly defined axioms according to precise rules, were already implicitly contained in the premises themselves, that were valuable only because they could create an order in the raw data that were difficult to achieve otherwise.

> Memex needs to graduate from its slavish following of discreet trails, even as modified by experience, and to incorporate a better way in which to examine and compare information it holds" (Bush 1959: 180).

One of the major achievements of Bush's vision was his firm belief that the technological evolution was based less on the presence of new technical devices, and "still more upon greater understanding of how to use them" (Bush 1959: 183). The central heritage of Bush could be represented by the centrality of the interactions between human beings and the machines. Technology can be revolutionary as far as it can be perceived and defined by the relationships with people and their needs. Hence, if Wiener was looking for the "human use of human beings" (1950), Bush was creating the intellectual and social space for the "human use of technology" which was the beginning of the tradition of

augmentation of human intelligence by the technical devices, instead of trying to simulate it via the machine.
Bush's influence could be detected directly or indirectly in the development of the computer as a communication device, as an augmentation for the human mind and in the studies on human-computer interactions. Some of the most influential scientists in the history of information and communication technologies, such as Joseph Licklider (the father of the intergalactic network, the ancestor of the Internet), Douglas Engelbart (the ideologue of the user-friendliness of computers), Ted Nelson (the inventor of the concept of Hypertext), Tim Berners-Lee (the inventor of the WWW) were directly or indirectly connected to Bush and to the Cybernetics tradition. These scientists contributed to the success of the computer as a powerful communication device, useful both for the other computers and for the human beings. The idea behind this project was described in details by Licklider, an experimental psychologist, expert in psycho-acoustic[260], who became fond of computers and suddenly had the insight of the new device as an active help in the scientific research. In 1960 he wrote a seminal paper on the *human-computer symbiosis* that was the official beginning of the metaphor of the computer as a 'social' tool that could communicate and interact with the users instead of being only programmed by them (Licklider 1960). The new idea of machine entailed that the device was an help, a facilitator, a connector of other people's work, and not a calculating machine. Licklider's machine could be considered partially outside the von Nueman's machine project, because his prototype took into account the role of the user and tried to focus not only on the data processing capabilities of the machine but on its versatility as a participant to a dialogue with the human being. The strength of the communicating machine metaphor changed deeply the device's potentialities: the new use of the machine had a profound impact on the perception of the device identity and on the understanding of the roles of the users. It was a revolutionary approach to the device that transformed irrevocably the information technologies.

Final philosophical observations

The machine built in the '40s of last century was at the confluence of various ideas that were not born at that time but came from far away. We can recognize in the debate around the intelligence of the machine the discussions the Seventeenth and Eighteenth centuries between the rationalistic and empiricist approaches to knowledge and its creation.

260 This area of research would be called neurosciences now.

Could the machine solve all problems by calculating the solution, as Leibniz would have suggested?
According to Gottfried Wilhelm Leibniz (1646-1716), one of the most prominent scientists, politicians and philosophers of his times, the best method to obtain certainty through knowledge was the creation of a system called *Characteristica Universalis* which would allow all the people who used it to "calculate" the solution for all the scientific and philosophical problems. The system consisted in two modules, one was the *lingua characteristica*, a sort of universal language that permitted to express in univocal form all the necessary and useful ideas in science or philosophy. The second module was called *calculus ratiocinator*, it was a method that allowed everybody to "deduce" via a calculus the correct conclusion for all the possible premises that were expressed correctly using the universal language. The use of this system, according to Leibniz, would avoid all possible mistakes and guarantee that all the conclusions were sound and true. The project was first envisaged when he was only 20 years old, but he kept on thinking of it all life long. In a letter to one of his many corresponded he declared:

> I am convinced more and more of the utility of this general science, and I see that very few people have understood its extent [...]. This characteristic consists of a certain script or language […] that perfectly represents the relationships between thoughts. The characters would be quite different from what has been imagined up to now. Because one has forgotten the principle that the characters of this script should serve invention and judgement as in algebra and arithmetic. This script will have great advantages; among others, there is one that seems particularly important to me. This is that it will be impossible to write, using these characters, chimerical notions […]. An ignoramus will not be able to use it or, in striving to do so, he himself will become erudite. (Letter to Jean Galloys December 1678, translated from French in Davis 2000: 16)

In this letter, he showed the major advantages of the new 'script' to his correspondent. First of all, it offered the guarantee that only the 'real' concept could be represented in it, and secondly it forbade ignorant people to use it, or alternatively they would become savant in the effort to master the method. Such a language would also allow the perfect correspondence of the relations among thoughts and it would also help the user to have clear and correct thoughts, adequate both to the external world and to the true consequences of all true axioms. All these results could be obtained by using a calculus similar to algebra or to arithmetic, which meant that once the notions were represented with the language symbols, it was very easy to 'calculate' the right conclusions. This project was only one of the long list of the dreams of reason by which human beings tried to control the access to knowledge and to guarantee the correctness of their conclusions. And the birth of the computer and of the consequent 'dream' of AI could be considered just another scene of the

same drama: the hope that certainty was achievable only by performing the right calculus.

On the other side of the epistemic range we have the work of David Hume (1711-1776), the philosopher who could be considered the champion of the empiricist tradition in the Eighteenth century. He discussed about knowledge and its characteristic in the first volume of the *Treatise of human nature*. Here, among other crucial questions, he stressed the central role of the association of ideas for knowledge creation, and declared:

> This uniting principle among ideas is not to be consider'd as an inseparabile connexion; for that has been already excluded from the imagination […] but we are only to regard it as a gentle force, which commonly prevails, and is the cause why, among other things, languages so nearly correspond to each other […]. The qualities, from which this association arises […] are three *viz.* RESEMBLANCE, CONTIGUITY in time or place, and CAUSE and EFFECT (Hume 1739: 10-11)

The basic characteristics of the association of ideas are the contingency of the connections and the central role of imagination in the creation of the links between ideas. Both these principles are central in Bush's description of the mind in its activity of accessing and connecting thoughts together. One of the consequences of the use of association of ideas in the communication machine paradigm was the introduction of *hypertextuality* as a new writing method that allowed to associate different ideas with each others, without following the linear flow of thoughts. The non-sequential writing model that was born in the human-computer interface research environment had a remarkable impact on the developments of the information technologies, whose consequences are still difficult to describe and foresee in details.

According to Michael Mahoney (2005), there are different communities that contributed to the machine design, and we would like to stress that there were also different philosophical approaches that merged together in the building of the device. This paper illustrates the need of the challenging research project of defining the different influences and professional mentalities that contributed to the computer, not only for the sake of the history of computing, but mainly for the understanding of the actual epistemological status of the information technologies studies. The digital age was not created just as an engineering affair, but stayed at the confluence of many professional influences and models. It is likely that the future results of these investigations could be surprising and the correct understanding of the multifarious nature of the computing machine could help in determining opportunities, risks, promising directions and threats in the future of computer science.

References

Aspray W. (1985), “The scientific conceptualization of information”, *Annals of the History of Computing*, 7:117-140.

Bush V. (1945/1999), “As we may think”, *The Atlantic Monthly*, July 176: 101–108, reprinted in Mayer P.A. (Ed.) (1999) *Computer Media and communication*, Oxford University Press, Oxford: 23-36, URL: http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush-all.shtml.

Bush V. (1959/1991), “Memex II”, in *Bush Papers*, MIT Archive, published in (Nyce and Kahn 1991), pp.165-184.

Copeland J. (2004) *The essential Turing*, Clarendon press, Oxford.

Cordeschi R. (2002), *The Discovery of the Artificial*, Kluwer Academic Publishers, Dordrecht.

Davis M. (2000), *The Universal Computer*, W.W. Norton & Co., New York.

Franchi S. and Güzeldere G. (2005), “Machination of the mind: Cybernetics and Aritificial Intelligence from automata to Cyborgs” in S. Franchi e G. Güzeldere (Eds.), *Mechanical bodies, computational minds*, Cambridge, MIT Press: 15-149.

Hume D. (1739) *A treatise of the human nature*, printed by John Noon, London.

Licklider J.C.R. (1960) “Man-computer symbiosis” in IEEE Transactions on human factors in Electronics, Vol. HFE-I, March 4–11; fully accessible online: http://memex.org/licklider.pdf.

Mahoney M. S. (2005), “The histories of computing(s)”, *Interdisciplinary Science Reviews*, vol.30, n.2, pp.119-135.

McCarthy J. (1959) “Programs with common sense”, in *Proceedings of a Symposium at NPL* 24-23 nov. 1958. London. H.M. Stationery Office. pp. 75-84.

McCarthy J., Minsky M.L., Rochester N. and Shannon C.E.(1955) *A proposal for the Darthmouth Summer research project on Artificial intelligence*, 31 August, http://www-formal.stanford.edu/jmc/history/darthmouth/darthmouth.html

McDermott D. (1987), “A critic of pure reason”, *Computational Intelligence*, Vol. 3. pp.151-160.

Neumann (von) J. (1945), *First draft of a report on the EDVAC*, (30 giugno), Contract, n. W-670-ORD-4926, Moore Scool of Electical Engineering, University of Pennsylvania; reprinted with corrections in *Ann. Hist. Comp*., 15: 25-75, (1993).

Numerico T. (2005), *Alan Turing e l’intelligenza delle macchine*, FrancoAngeli, Milano.

Shannon C. (1963), “Computers and automation – progress and promise in the twentieth century” in Higginbotham S.W. (Ed.) *Man, science, learning, education*, the semicentennial Lectures at Rice University, Rice University Studies, Vol. XLIX, Suppl. 2: 201-211.

Turing A.M. (1937), “On Computable Numbers, with an application to the Entscheidungsproblem”, Proc. Lond. Math. Soc. (2) 42 pp 230-265 (1936); correction ibid. 43, pp 544-546 (1937), reprinted in Copeland 2004: 58-90.

Turing A.M. (1945), “Proposal for the development in the Mathematical Division of an Automatic computing engine (ACE)”, Report to the Executive Committee of the National Physical Laboratory del 1945, in D. C. Ince (Ed), (1992) *Collected Works of A. M. Turing: mechanical intelligence*, North-Holland, Amsterdam: 1-86.

Turing A.M. (1947), “Lecture to the London Mathematical Society on 20 February 1947”, in Copeland 2004: 378-394.

Turing A.M. (1948), “Intelligent Machinery” Report, National Physics Laboratory, in B. Meltzer D. Michie (Eds) *Machine intelligence*, 5 (Edinburgh Univ. Press, 1969); reprinted in Copeland 2004: 410-432.

Wiener N. (1950), *The Human Use of Human Beings*, Houghton Mifflin, Boston.

COMPUTATIONAL EPISTEMOLOGY AND E-SCIENCE.

A NEW WAY OF THINKING.

Jordi Vallverdú i Segura

Philosophy Department, E-08193

Universitat Autònoma de Barcelona

Tel. 00 34 93 581 16 18; Fax. 00 34 93 591 20 01

jordi.vallverdu@uab.es

Abstract: Recent trends towards an e-Science offer us the opportunity to think about the specific epistemological changes created by computational empowerment in scientific practices. In fact, we can say that a computational epistemology exists that requires our attention. By 'computational epistemology' I mean the computational processes implied or required to achieve human knowledge. In that category we can include AI, supercomputers, expert systems, distributed computation, imaging technologies, virtual instruments, middleware, robotics, grids or databases. Although several authors talk about the extended mind and computational extensions of the human body, most of these proposals don't analyze the deep epistemological implications of computer empowerment in scientific practices. At the same time, we must identify the principal concept for e-Science: *Information*. Why should we think about a new epistemology for e-Science? Because several processes exist around scientific information that require a good epistemological model to be understood.

Keywords: e-Science, epistemology, computation, extended mind.

Introduction

In the middle of XXth Century a new scientific phenomenon appeared, that is, *Big Science*. The Manhattan project for the creation of an atomic bomb, the Apollo mission to the Moon, NASA's Voyager project for planetary exploration on a grand scale, particle accelerators and the Human Genome Project are different and consecutive historical examples of that process (Capshew & Rader, 1992). Their scientific, budgetary, and technological immensity make these research projects archetypical big science.

The same century experienced another important occurrence: the development of electronic computer machines. Following on from the seminal ideas of Alan Turing and John von Neumann, several huge computer machines were created and employed initially for military uses, but the transistor and microprocessor revolution enabled the creation of microcomputers, facilitating the implementation of computers in all kind of situations. Finally, the communication revolution of satellite technologies and the development of the Internet connected all these machines together and enabled a new way of life and thinking.

Thus, science has turned into e-Science, that is, computationally intensive science. This new kind of science is also the type of science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing. Recent trends towards an e-Science offer us the opportunity to think about the specific epistemological changes created by computational empowerment in scientific practices. In fact, we can say that a computational epistemology exists that requires our attention. By 'computational epistemology' I mean the computational processes implied or required to achieve human knowledge. In that category we can include AI, supercomputers, expert systems, distributed computation, visualization and imaging technologies, virtual instruments, middleware, robotics, grids or databases.

Although several authors talk about the extended mind and computational extensions of the human body (Clark & Chalmers, 1998; Humphreys, 2004; Hutchins, 1995; Norman, 1997), most of these proposals don't analyze the deep epistemological implications of computer empowerment in scientific practices. They talk about new human physical and mental environments, not about new ways of reasoning, in the broader sense of the term.

At the same time, we must identify the principal concept for e-Science: *Information*. Sociologists like Castells (1996) or philosophers like Floridi (2001) speak respectively of a Network Society with a 'culture of real virtuality', an open space sustained by the Information Technology (IT) revolution (and changes inside capitalist economic models and the

pressure of new cultural and social movements), or a new space for thinking and debating, the *infosphere* (Floridi, 2001). We could also talk about a Philosophy of Information (Floridi, 2002, 2003). My point of view about all the processes in which scientific information is involved in e-Science can be condensed in the next table:

Processes	Facts
Creation/Discovery	Data Tsunami: Petabytes of data Virtual Instruments Ontologies Artificial Intelligence and Expert Systems
Management: search-access-movement-manipulation-mining	Databases: Complex, hierarchical, dynamic. Software Middleware
Understanding	Computerized Modelization Imaging Information integration
Evaluation	Computational, open
Communication	Electronic open access journals: PloS.
Work strategies	Delocalized Network Cooperative Dynamical Interoperativity Spatially distributed cognition
Funding	Symbiosis between Public & Private (HGP-Celera, Roslin-PPT…)
Control	Beyond national control

Now, we are faced with a new science, e-Science, highly related to computers, the Internet and new ways of information processing. We should now think about its epistemological roots.

1. e-Science & epistemology.

We must admit that despite the fact that there have been several philosophers who have tried to show the radical implications of computation in human reasoning (Bynum & Moor, 1998; Thagard, 1988; Landow, 1994; Rouet, 1996; Mitcham, 1994; Vallverdú, 2005), this has not implied the design of a new epistemology for e-Science.

Why should we think about a new epistemology for e-Science? Because several processes exist around scientific information that require a good epistemological model in order to be understood. We face new problems like: (1) reproducibility of experiments made with virtual instruments and

remote databases, and the evaluation of the algorithms implied... (Knight & Leveson, 1986); (2) classical statistical dilemma implied in contemporary calculus: Bayesian Networks vs. Frequentism, with the spectacular increment of Bayesian models due to computational facilities; (3) cognitive biases in imaging technologies: the evaluation of virtual models. (Giere, 2003; Rapp. 2003, Latour, 1986). For example, visualization as 'human distortion of data' (Humphreys, 2004); (4) distributed Computing and coordination of exponential functions (small differences between hundreds of personal computers can offer false scientific results); (5) divergences in the designed middleware, useful to work with different databases; (6) evaluation and communication through new open access journals; (7) the AI tools introduced to work with petabytes of data. Just as an example from the field of physics: the Large Hadron Collider (CERN, Genova) produces **15 petabytes/year**. To be able to analyze such an amount of data the LHC Computer Grid was created (http://lcg.web.cern.ch/LCG); (8) the evaluation of complex and non-evident computational processes (Norman, 1997); (9) automated discovery (Langley, 2000; Valdés-Pérez, 1999), or (10) e-learning through e-Science.

These are just some of the problems that we can find in the new e-Science that require a new epistemological approach. Extended Cognition, Philosophy of Information, Cognitive Sciences and Artificial Intelligence approaches can help by working together to define a new epistemological model for e-Science.

2. Science and computers.

From the very beginnings of human sedentary culture, the complexity of day-to-day necessities required a way to remember and compute amounts of data. Marks on wood or bone were the first way to remember large numbers. Ancient humans also used their bodies, hands or fingers to count and to calculate. When the numbers became too large, it was necessary to design systems to calculate them: do arithmetic on clay tablets, papyrus, leather or paper. But special machines were also created to perform those calculations: the abacus. There subsequently appeared the Antikythera Device, Napier's Bones, the Pascalene, the Leibniz's 'Stepped Reckoner', Babbage's Analytical Engine, Hollerith's Electric Tabulating System, Bush's Differential Analyzer, Zuse's Z1, Z2 and Z3, Colossus, ENIAC… different kinds of progressively hand, mechanical, electromechanical and electronic computing machines.

Beyond practical accounting, scientific disciplines required people as well as machines to perform those calculations. Astronomy was one of the disciplines that required those machines and specialists although one of the main fields of numerical activity has been ballistics.

During the early 1940's, approximately 75 young female mathematicians were employed as a "computer" by the University of Pennsylvania's Moore School of Engineering. These women were responsible for making calculations for firing tables and bombing trajectories, as part of the war effort. The need to perform the calculations more quickly prompted the development of the ENIAC, the world's first electronic digital computer, in 1946[261].
From the second half of 20th Century, computers have been employed in all the academic disciplines: medicine, archeology, mathematics, physics, logics, astronomy, anthropology … So, we have computationally empowered science. This fact has involved deep changes in the way we understand and carry out scientific practices. Just as an example, we can look at mathematics and the concept of 'proof'. Before computers, mathematical proofs were a formal series of statements showing that if one thing is true something else necessarily follows from it. Or, more precisely, a proof is a demonstration that, assuming certain axioms, some statement is necessarily true. There are several proof techniques like: direct proof, proof by induction, proof by contradictions, proof by construction, proof by exhaustion, probabilistic proof or combinatorial proof.
Actual computer-aided proofs are considered the result of brute force, inelegant and *ugly* proofs. The first time a computer was utilized to prove a theorem was in 1976 with the Four-Color Problem. More than twenty years later, Kepler's conjecture was also solved with the aid of highly intensive computer work (Wilson 2003; Szpiro 2003). The general strategy consisted in reducing a proof to a finite list of possible counterexamples, and then eliminating them one by one. What is the problem with this? We cannot verify every step of the procedure. In Thomas Hales demonstration of Kepler's conjecture, the 12 reviewers of *Annals of Mathematics* worked for several years until they decided that the proof *seemed* true, although they could not certify it. The reviewed materials of the full proof contained over 250 pages, with 3GB of computer data files (computer code, data files for combinatorics, interval arithmetic and linear programs).

The use of computer simulations in astronomy, physics, biology or toxicology, are also the demonstration of the extended use of these machines in scientific research. But, strictly speaking, these are not cases of e-Science, but cases of contemporary science in which computational tools are embedded in research. From my point of view this is a stage previous to e-Science in the evolution of scientific dynamics. Similarly, when the Second World War led to the appearance

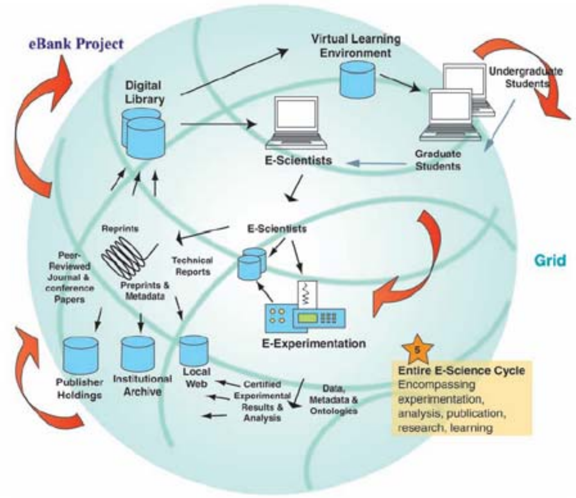
261 http://www.cs.yale.edu/homes/tap/past-women-cs.html

of Big Science, not all disciplines grew up to embrace the new kind of big infrastructures. However the main tendency oriented the evolution of international projects and quantification of evaluation.

3. e-Science and computers.

At the beginning of 21st century, a new kind of science has appeared, namely, e-Science. This computationally intensive science has its own dynamics, based on the intensive use of computers and information technologies. It has received several names, such as ‘cyberscience’, ‘robot science’, ‘automated science’ or ‘virtual science’, but ‘e-Science’ is perhaps the most successful denomination. It can also be considered as the *combination* of three developments: (a) large-scale computing resources; (b) access to massive, distributed and heterogeneous datasets, and (c) use of digital platforms for collaboration and communication. We can look at the whole process with this diagram, from *Science*, vol. 308, p. 820

:



With e-Science, we are not talking about normal science made with computers, but a completely new way to perform scientific activities, from research to evaluation, communication or education. We can affirm that there is a new culture of scientific activity. The influence, for example, of open source communities has created a breakdown in the communication system based on classic scientific journals. The whole publishing process is now considered to be unsatisfactory by the new generation of scientists as: authors pay to be published, reviewers work is unpaid (although it takes them time) and journal subscriptions are very expensive. So, an author receives no money for her/his publication but must pay to consult it. It is a nonsensical process that can be changed with open access journals like PLoS or arXiv, with an open and free attitude towards research. At the same time, these journals don’t receive industry pressure, as has happened and has been denounced by the main biomedical journals. In September 2001, about 12 of the world's most prominent medical journals (Annals of Internal Medicine, the Journal of the American Medical Association, the New England Journal of Medicine, the Canadian Medical Association Journal, the Journal of the Danish Medical Association, the Lancet, MEDLINE/Index Medicus -a medical database-, the New Zealand Medical Journal, the Journal of the Norwegian Medical Association, the Dutch Journal of Medicine and, finally, the Medical Journal of Australia and the Western Journal of Medicine) issued a joint editorial stating that they would reject any scientific studies that do not come with an assurance that the sponsor -- whether a drug company or another organization -- gave researchers complete access to the data and freedom to report the findings. It was a response to what editors said was excessive control by drug companies over how the results of studies they sponsored were analyzed, interpreted and reported.

But open access journals are also leading to a deep change in peer review systems. As an example, the proof of Poincaré Conjecture published in arXiv (www.arxiv.org) by Grisha Perelman, is not peer reviewed although he has been offered to be published by *The Journal of Geometric Analysis*. Are we at the end of the classic peer review system and at the beginning of a new way to communicate and validate scientific information?

In 1942 Robert K. Merton explained[262] the four 'mertonian norms' of scientists desired behavior: *communalism* - science is an open community; *universalism* - science does not discriminate; *disinterestedness* - science favors an outward objectivity; and, *organized skepticism* - all ideas must be tested and are subject to community scrutiny. His ideas had great success and were received with open arms by science theoreticians, at a historical moment, the Cold War, at which time a confrontation between 'democratic' and 'communist' approaches to science analysis and development existed. Some years later, John Ziman added *originality* to Merton's four norms and reformulated them according to the acronym CUDOS[263].

The idea of communalism was very conflictive, because of the possible conceptual ties with communism. But in the end, we must remember that the real confrontation wasn't between democracy and communism, but between capitalism and communism. And we have seen previously the threatening forces of private industry over scientific journals and the reaction of the scientific community towards open access journals. But there are also critics who are against these open source trends, as we can see in a contemporary essay by Jaron Lanier **"**Digital Maoism: The Hazards of the New Online Collectivism"[264]. To be honest, nothing is black and white, and we must consider social forces inside scientific dynamics.

Open source culture has changed the meaning of several forms of thinking about scientific practices. And not only communication processes with open access journals, but also the software necessary for developing the research. For example, BOINC[265] is free, open-source software for volunteer computing and desktop grid computing. You can use the idle time on your computer (Windows, Mac, or Linux) to do all sorts of scientific research. A partial list of current projects which use BOINC: Malariacontrol.net, SETI@home, Climateprediction.net, SIMAP,

---

262 "Science and Democratic Social Structure" in Social Theory and Social Structure – Enlarged Edition, New York: The Free Press, 1968. The article was first published in 1942 as "A Note on Science and Democracy", Journal of Legal and Political Sociology 1: 115-126.

263 Reliable Knowledge, Cambridge University Press, 1978. P.6-8

264 http://www.edge.org/3rd_culture/lanier06/lanier06_index.html.

265 http://boinc.berkeley.edu/.

World Community Grid, SZTAKI Desktop Grid, LHC@home, Quantum Monte Carlo at Home, BBC Climate Change Experiment, Einstein@home, Tanpaku, Rosetta@home, Seasonal Attribution Project, Predictor@home. So, ***distributed computing*** can be considered as the introduction of civil society inside scientific practices, and an open knowledge construction process.

Supercomputing[266] environments also have an important role to play in developing advanced research in several scientific fields. In my country, Catalonia, we have one of the most powerful supercomputers in the world, the Mare Nostrum, in the BSC (http://www.bsc.es). Taken as an example of other supercomputing infrastructures, the BSC has the following objectives: (1) investigate deep computing, computer architecture and Information Technology in general, (2) collaborate in e-Science research projects with well-known international researchers, (3) manage the center resources in order to help researchers and scientists make the best use of the supercomputer technology, (4) develop innovative solutions in collaboration with private companies, (5) inform society about the benefits of Information Technologies and, (6) provide training to expert professionals in the different areas of research. A major tool, for dealing with important necessities. The paradigmatical place occupied formerly by particle accelerators in Big Science, now belongs to supercomputers in (computational) science. They are the symbol of a new era and politics of research.

Grids and middleware. Computational Grids enable the sharing, selection, and aggregation of a wide variety of geographically distributed computational resources (such as supercomputers, compute clusters, storage systems, data sources, instruments, people) and presents them as a single, unified resource for solving large-scale compute and data intensive computing applications (e.g. molecular modeling for drug design, brain activity analysis, and high energy physics). These grids can also show several functioning gaps, such as gaps in Security, Workflow, Notification Service, Meta-data and Semantic Grid, Information Grid Technology, Compute/File Grids, Grid Technology, Portals and Problem Solving Environments, Grid-Network Interface, Education and Support Gaps[267].

All these facts, open source materials, distributed computing, supercomputing and grids are part of a common project of e-Science, a new kind of science, with new problems and with its own epistemological analysis. Knowledge is now a process of intensive interactions between human and machines, not only at an observational level but also at a cognitive (visualizations, storing, calculations, and data mining…).

---

266 http://www.top500.org/.

267 www.grid2002.org/ukescience/gapresources/GapAnalysis30June03.pdf.

4. Computational epistemology.

In 1991, World Scientific Publishing printed a book in Singapore from a Hungarian scientist, Tibor Vámos. Its title, *Computer epistemology*, advanced some of the ideas discussed here: the complexity of recent human knowledge developed by utilizing computational tools requires a new epistemological approach. Old ideas are converted into new ones, under the same names but with different meanings or extended computational environments: "model", "uncertainty", "logic", "learning" or "proof", are some of them. Machines are producing knowledge. Genetic algorithms have discovered laws of nature (without proving them or explaining why they are true). Expert systems discover new conceptual relationships between our information and the world (Quinlan 1979), that is, they create new knowledge which would have been otherwise hidden from us. In the way that without telescopes our eyes could never have seen Jupiter's moons, our minds could never have reached some information from the huge amounts of raw data. So, if contemporary science is based on computer processes, we must have an epistemology of computing.

One of the most important problems of bioinformatics is the one of the reproducibility of computational results. In the classic models of understanding of scientific activity, it has been considered, since the Renaissance, that science is something experimental and that its explanatory success and predictive value is related to the possibility of reproducing the experiments which demonstrate a hypothesis or theory. But, due to the nature of computational environments (usually as grids), it is very difficult to reproduce exactly a virtual experiment. There are several reasons: the use of non-unified standards for data storage, diverse processing algorithms...

We can also find cases of falsifiability in computational science, such as the famous experiment of Knight & Leveson (1986). That experiment analyzed the failure probabilities of multi-version programs. Conventional theory predicted that the failure probability of a multi-version program was the product of the failure probabilities of the individual versions. However, John Knight and Nancy Leveson observed that real multi-version programs had significantly higher failure probabilities. In essence, the experiment falsified the basic assumption of the conventional theory, namely that faults in program versions are statistically independent[268].

268 Tichy (1998): 33.

As an example of possible problems using computational devices we can find these errors:

(a) Human errors, or GIGO[269]:
    a. Model programming/design: *bugs.*
    b. Data introduction.
    c. Data evaluation (also due to problems in the visualization processes).

(b) Machine errors
    a. Software:
        i. Introduced data.
        ii. Coordination and verification: static (font code) or dynamic (using the program).
        iii. In the middleware of grids: provenance.
    b. Hardware:
        i. Cosmic radiations: a supercomputer such as the Cray-1A undergoes an undetected error per thousand hours of operation, and this occurs through a random change of a *bit* in computer memory, brought about by, of all things, cosmic radiation.
        ii. Those of floating-point arithmetic controlled in 1985 by the *IEEE 754* standard.
        iii. Errors in microprocessor design. The famous Intel 'Pentium' microprocessor was an outstanding case. Discovered in 1994, the error consisted in the fact that the Pentium microprocessor was not able to solve certain calculations of floating comma under certain conditions of calculation. We can affirm that the problem should not have happened under the standard IEEE 754, described previously.

All this leads us to question the validity of the data that we obtain through our computers without it being univocally related with the calibration of the instruments, but rather with the totality of the cognitive process which allows the creation of new knowledge. This is really important, since in many scientific disciplines, we can only think about and discover the secrets of nature through these complex instruments. Our mind has extended towards the computers, which is the reason why

---

269 Garbage In, Garbage Out (abbreviated to GIGO) is an aphorism in the field of computer science. It refers to the fact that computers, unlike humans, will unquestioningly process the most nonsensical of input data and produce nonsensical output. It was most popular in the early days of computing, but has fallen out of use as programs have become more sophisticated and now usually have checks built in to reject improper input. Font: http://en.wikipedia.org/wiki/Garbage_in,_garbage_out.

our epistemology also depends on them. ‘Knowing’ no longer constitutes the strict scope of the human mind (Clark, 2003).The automatization of the problem of the four colors or the conjecture of Kepler, both already theorems, are examples of this new form of obtaining knowledge.

5. Conclusions.

I propose an Integrative Approach to Computation (IAC) which could enable the development of a comprehensive computational epistemology, following these main points:

- ✓ Unify data status (natural-observational, experimental, and computational) under a basic unit: **scientific information (si).**
- ✓ **Unify working ontologies** (CBL, XML, OAGIS, OCF, OFX, RETML, UN/SPSC...). Kinds of ontologies: domain, task, quality, value… for a *better interoperability.*
- ✓ **Make metaproof** decisions about the values of proofs: proofs by Modeling, Simulation or Experiment. So, are real experiments superfluous?).
- ✓ **A dynamic** model of e-Science based on information (creation-analysis, communication): more agents (i.e. civil society in distributed computing).
- ✓ A New epistemology of (computational) instruments, that is, a **synthetic epistemology,** where instruments are **active** (instead of passive) knowledge creators.
- ✓ **Develop a transcognitive model**: ‘extended mind’ beyond human-centered model “robots-AI-human”. AI & expert systems are not *just* extensions of us.
- ✓ **Consider open source** as collaborative knowledge construction (peer-to-peer grids & distributed computing), in contrast to **private science**.

This is an ongoing project I am developing at my university. This research has been developed under the main activities of the TECNOCOG research group (UAB) about Cognition and Technological Environments, [HUM2005-01552], funded by MEC (Spain).

Bibliography

Boden, Margaret A. (ed.) (1990) The philosophy of artificial intelligence, UK: Oxford University Press.

Bynum, T.W. & Moor, J.H. (1998) The Digital Phoenix. How Computers Are *Changing Philosophy,* UK: Blackwell Publishers.

Capshew, J.H. & Rader, K.A. (1992) "Big Science: Price to the Present," *Osiris*, 2(7): 14-22.

Castells, Manuel (1996) *The Rise of the Network Society*, Oxford: Blackwell.

Clark, A. & Chalmers, D. (1998) “Embodied, situated and distributed cognition”, *in* W. Bechtel & G. Graham (eds.) *A* Companion to Cognitive Science, Malden (MA): Backwell.

Floridi, Luciano (2001) “Information Ethics: An Environmental Approach to the Digital Divide”, *Philosophy in the Contemporary World*, Volume 9 (1: Spring-Summer).

Floridi, Luciano (2002) "What is the philosophy of information?", *Metaphilosophy*, 33(1-2): 123-45.

Floridi, Luciano (2003), "Two approaches to the philosophy of information", Minds and Machines, 13(4): 459-69.

Giere, R. (2003) The role of computation in scientific cognition. *Journal of* Experimental & Theoretical Artificial Intelligence, 15, 195-202

Humphreys, Paul (2004) Extending Ourselves. Computational Science, Empiricism and Scientific Method, USA: OUP.

Hutchins, E. (1995). *Cognition in the Wild.* MIT Press.

Knight, J.C. & Leveson, N.G. (1986) "An Experimental Evaluation of the Assumption of Independence in Multi-version Programming", *IEEE Trans. Software Eng.*, Jan: 96-109.

Landow, George P. (ed.) (1994) *Hyper/Text/Theory*, USA: Johns Hopkins University Press.

Langley, P (2000) "The computational support of scientific discovery", International Journal of Human-Computer Studies, 53(3): 393-410.

Latour, B. (1986) "Visualization and Cognition: Thinking with Eyes and Hands", Knowledge and Society, 6: 1–40.

Mitcham, Carl (1994) Thinking Through Technology. The Path between *Engineering and Philosophy,* Chicago: Chicago University Press.

Norman, D.A. (1997) *The Invisible Computer*, Cambridge, Massachusetts, MIT Press.

Quinlan, J. R. (1979) "Discovering Rules from Large Collections of Examples: a Case Study", in Michie, D. (ed.), Expert Systems in the Micro-Electronic Age,Edinburgh: Edinburgh University Press, pp. 168-201.

Rapp, David N. *et al* (2003) "The impact of digital libraries on cognitive processes: psychological issues of hypermedia", Computers in human behavior, 19: 609-628.

Rouet, Jean-François (1996) *Hypertext and Cognition*, USA: Lawrence Erlbaum Associates, Inc.

Szpiro, G.G. (2003) Kepler's Conjecture. How some of the greatest minds in history helped solve one of the oldest math problems in the world, USA: John Wiley & Sons.

Thagard, Paul (1988) Computational Philosophy of Science, Cambridge (MA): MIT Press.

Tichy, W.F. (1998, May) "Should Computer Scientists Experiment More?", *Computer*, 32-40.

Valdés-Pérez, Raúl E. (1999), "Principles of human-computer collaboration for knowledge discovery in science", *Artificial Intelligence*, 107: 335-246.

Vallverdú, Jordi (2005) "Hypertextual Thoughts", *Revista Portuguesa de Filosofia*, In Press.

Wilson, R. (2003), Four Colors Suffice: How the Map Problem Was Solved, USA: Princeton University Press.

IMPLEMENTING NEGATIVE MORAL COMMANDS

INVESTIGATING EXPLICIT ETHICAL AGENTS

V. WIEGEL

Delft University of Technology, Faculty of Policy, Technology and Management

PO Box 5, 2600 AA Delft, The Netherlands

v.wiegel@tbm.tudelft.nl

1. Introduction

Various authors such as Floridi, Van den Hoven, Moor [10] have proposed classifications of agents[270] with increasing moral reasoning capacity and moral relevance. For the purpose of this article I will follow [10] distinguishing four categories of ethical agents: ethical impact agents, implicit ethical agents, explicit ethical agents and full ethical agents. Ethical impact agents are all agents that have by their very nature and existence an ethical impact. The ethical aspect is not 'in' the agent but in the influence on their environment. Implicit ethical agents have moral considerations designed and built into them. But they cannot be said to reason about the moral aspects. Their make-up is such that they simply cannot violate particular moral rules. Explicit ethical agents are

270 By an agent I will refer to both artificial and human entities in a loose sense that includes also machines and virtual objects.

agents that can make moral judgements and provide some account of how they arrived at their judgements. Full ethical agents are agents that are engaged in making ethical judgements relating to complex, new situations with the ability to provide a plausible justification. A full ethical agent "lives a moral life". They have a free will, intentionality and conciousness.

Implementing moral reasoning in artificial agents is a very broad and complex topic. It is important to stress that the intention is not to construct anything that might lay a claim to getting close to human moral reasoning. Both the current technology and our understanding of the moral discourse are far too limited to even consider embarking on such a venture. This is not to say that it is impossible to create an artificial moral agent one day that can compare to humans in its moral reasoning. But definitely not yet. My sentiment in this respect is the same as Wooldridge when working on his logic for artificial, rational agents.

> "Belief, desire and intention are in reality far too subtle, intricate and fuzzy to be captured completely in a logic [...] if such a theory was our goal, then the formalism would fail to satisfy it. However, the logic is emphatically *not* intended to serve as such a theory. Indeed, it seems that any theory which *did* fully capture all nuances of belief, desire and intention in humans would be of curiosity value only: it would in all likelihood be too complex and involved to be of much use for anything, let alone for building artificial agents."
>
> Wooldridge, [15:91]

Hence the focus in this article is on explicit ethical agents because they have some degree of autonomy, and provide a challenge to our understanding but are yet within the realm of the possible in the years to come.

2. Implementation

Implementation is done in three stages: modelling, design and coding. The first step is the modelling of the required behaviour. For this purpose DEAL (deontic epistemic action logic) is used in conjunction with the BDI (belief desire intention) model. These models consist of standard modal logic operators. They are used as specification language. This provides a language to capture requirements that is stricter than our everyday language but more relaxed than the logic reasoning with axiomatizing

and theorem proving. For the purpose of this article I will only provide an overview of the operators. For a more detailed discussion and examples the reader is referred to Sergot[12], Wiegel[13].
Reasoning about what one knows or believes is captured by epistemic logic, which has two operators: Bi (agent i believes that) and Ki (agents i knows that). KiΦ states that agent i knows that Φ[271].
Action logic is a branch of modal logic. Its operator is STIT, “see to it that”.

1. [i STIT: Φ] means agent 'i' sees to it that 'Φ' is done or brought about.

Predicate logic assigns predicates to actions and situations.

2. *G*(Φ) means 'Φ' is 'G'

G can be interpreted as morally good and 'Φ' as a situation or an action that brings about some situation. Combining the above we can write

3. *G*([i STIT Φ])

if an act is morally good – the act is good but the outcome might or might not be good.
Deontic logic has one basic operator,

4. O(Φ) it is obligatory that Φ

Two other operators can be derived from this primitive operator:

5. P(Φ)

it is permissible that Φ, or alternatively ¬O¬Φ, and

6. F(Φ)

it is forbidden that Φ, or alternatively O¬Φ), [8:284].
Following Wooldridge's definitions [15][272]

7. (Bel i Φ) means i believes Φ
8. (Int i Φ) means i intends Φ
9. (Des i Φ) means i desires Φ

These elements can be combined to construct moral propositions. Consider the following proposition, which is for demonstration purposes only and not necessarily true or a desirable property of an artificial agent.

10. $Bi(G(\Phi)) \rightarrow O([i \text{ STIT } \Phi])$

meaning if i believes that 'Φ' is morally good than i should act in such as way that 'Φ' is brought about[273].

To create support for the above modelling components I will use the following implementation elements from the JACK development environment [1].

- Beliefsets – beliefs representing the epistemic dimension
- Events – goals and desires, for the goal-directed behaviour
- Actions, plans and reasoning methods – representing the intentions and action logic
- Agent – the container for the other elements
- Java programming language[274]

The deontic dimension is a complex dimension build up from the above elements.
Beliefs represent the agents view of its outer world. Beliefs are implemented as a first-order, relational model, called beliefset. Each beliefset has

- zero, one or more key fields[275] (all usual data types plus a logical member)

---

271 In the remainder I will use only the Belief operator as this fits with the choice for the BDI-model.

272 Wooldridge designed an extensive and impressive logic for rational agents with many more components such as for example path connectives and quantifiers. For the current purposes the BDI operators suffice.

273 All the above provide still a relatively simple moral propositions. Subsequent developments will increase the complexity and problems of agglomeration will surface. Mechanisms for reasoning about and decision on conflicting obligations are addressed in subsequent sections.

274 In fact the other elements are an abstraction layer on top of the Java programming language. The development are not limited to these elements but can be extended with specific Java code.

- one or more value fields,
- a set of queries[276]

The logical consistency of a beliefset is maintained automatically[277]. Beliefsets can be modelled using 'open world' and 'closed world' semantics. In closed world semantics something is either true or false. The open world semantics allows something to be unknown. In the implementation the closed world beliefsets contain only tuples that are true. Tuples that are not stored are assumed false. In open world semantics both true and false tuples are stored. Tuples not stored are assumed unknown.

Desires and goals are what drives an agent in the JACK agent environment. They give it its goal-directed behaviour that allows it to reason about what it wants to achieve independently of how (which is done through actions / plans). It also allows for pro-active rather than reactive behaviour. Desires, as represented by `BDIGoalEvents`[278], are a special type of events. Events can be inter-agent or intra-agent. The former represent the usual interaction between entities, the exchange of information, requests and answers. The latter represents fine-grained internal reasoning processes. Events can be posted (for internal processing only) or send (for inter-agent communication) in various ways (polymorphy[279]). A BDI event can potentially be handled by multiple plans. When there are multiple applicable plans another event, the PlanChoice event, can be raised which is handled in turn by a meta-level plan (see plans below). Events can be posted or sent by agents from within plans, by external components (other programs), and by beliefsets

---

275 A key field is a data field that is used for indexing and that can be used to find particular beliefsets.

276 There are various kinds of queries: linear; indexed; complex – combining simple queries; function – developer coded, special queries. How they work is not important in this context. It serves to show that a fine granular set of mechanisms is available to unlock information.

277 If a new fact is added that contradicts an existing one the old state will be knocked. To allow sophisticated reasoning (the agent may want to 'think' before really knocking an existing belief) events can be posted. Beliefsets can post events in case 1) new facts are to be added, 2) have been added, 3) the state of a belief changes (e.g. true to false) when new facts are added or removed due to negation or key constraints, 4) beliefs are removed. In response to the event the agent can decide whether to accept the change in the beliefset.

278 BDIGoalEvent is the term used in JACK to denote a special type of events that can be used for complex (meta-level) reasoning. Normal events in contrast are just sent and processed by the appropriate plan but does not allow for explicit meta-level reasoning.

279 Polymorphy refers to the possibility to have multiple implementations of the same basic function each with different inputs (signature in software engineering terms). The main benefit is that it allows for rich, fine granular behaviour.

An agent has one or more plans at its disposal to achieve its goals. A plan is a sequence of atomic acts that an agent can take in response to an event. Committing to a plan, choosing a plan is like forming an intention. There are potentially several plans that can handle an event, and each plan can handle only one type of event. In order to determine which plan will handle an event (if any) there are two methods: `relevance()` and `context()`. The `relevance()` method determines which instances (all or some) of an event type can be handled. An event can carry various information which allows the `relevance()` method to determine whether or not to handle the event. From all relevant plans the `context()` method determines next which are applicable. The context method is a logical expression that tries to bind the plan logical members[280]. For each binding a plan instance will be created. E.g. an agent might have a plan to help some other agent in need. But it will only help agents from the same tribe, which is determined through the `relevance()` method in conjunction with information contained in the event message member. Next, it tries to bind a logical member AgentsInNeed against a beliefset containing all tribe agents, and an indication whether they are in need. For each of the bindings (tribe agent) a plan instance will be created. It might execute one plan, all plans till the first succeeds, or all plans.

A plan can have some meta information associated to it – accessible through `PlanInstanceInfo()`. This can be a ranking number that can be given a cardinal or ordinal interpretation. This information can be used to reason at a meta-level in case there are multiple, applicable plans. In that case a special event, `PlanChoice` event, is raised. This event can be handled by a meta-level plan that facilitates reasoning about the various courses of course, the precedence of one over the other.

Explicit, meta-level reasoning is the finest granular reasoning facility. But plans also have a prominence, that is the order in which they appear in the agent's make up. If no other ordering information is provided plans will be executed according to their prominence. Finer ordering can be achieved through precedence, providing a ranking to a plan which can be accessed through the `PlanInstanceInfo()` method.

When chosen for execution the body of the plan is executed. This is the core element of the plan that contains the detailed instructions (statements) of the plan. This is made up by the Java programming language and extended with the JACK reasoning method statements. The reasoning method statements are special JACK agent language

280 A member can be thought of as an attribute. A logical member is like normal data member, such as a string or an integer, but with the addition of following the rules of logic programming. Binding is the process of finding values for the members that match the logical conditions.

constructs that facilitate the control over reasoning, and specific agent behaviour. These statements are implemented as finite state machine[281].

3. Implementing negative moral commands

Using the equipment outlined in the preceding section I now discuss the way in which these concepts and ideas can be implemented in software[282]. In this section I will first show how the basic elements (operators) are constructed. Using these basic elements more complicated propositions (pairwise combinations of modal logic operators) will be created.

3.1. Belief

A belief that a state Φ is morally good can be implemented as a tuple describing that state in a beliefset of morally good states[283]. I will model two basic beliefsets: one representing the moral obligations, and one containing the specific actions or states with their deontic status[284]. I will model both under open world semantics. This requires the designer to be specific and as complete as possible. E.g. modelling obligations under closed world semantics will state as morally wrong everything the modeller forgot to specify as morally good. It also reflects better the fact the moral agents are not omniscient. Dealing with uncertainty is also an important aspect of moral behaviour.

a) Listing beliefset MoralObligations

```
public beliefset MoralObligations extends
OpenWorld {
     #key field String strObligationName
     #key field String strSphere
     #value field String strMoralProposition
```

281 A finite state machine is an execution model in which the execution of a step cannot be stopped but must be completed before anything else can be done.

282 In this section the software components and the logical operators are used alternatingly. I will often when using one append the other in brackets to make clear what is being referred to.

283 There is another option: implementing moral beliefs as a tuple describing that state plus its moral evaluation in a beliefset of all states. This option decreases the redundancy of information stored since all relevant aspects of a state are in one beliefset. When the moral dimension, and possible other dimensions, are stored in separate beliefsets the key fields need to be stored in all beliefsets increasing the redundancy. On the other hand, storing all aspects in one large beliefset increases the overhead of maintaining and querying that beliefset. It has a negative impact on the processing performance. I cannot say that one or the other might be closer to the way human cognition functions. The option I use offers greater clarity in representation.

284 This status is derived at run time and not before.

```
        #value field String strText
        #value field String strType
        ....
}
```

The field `strObligationName` will contain a reference to the obligation[285]. The `strSphere` field is to recognize that obligations might have a restricted application domain, sphere. `strMoralProposition` contains the logical proposition representing the moral obligation, and `strText` its textual description. The `strType` is meant to be able to distinguish between states and actions. Beliefsets automatically have a boolean indicator signifying truth or falseness of the tuple.
The second beliefset contains tuples with statements about concrete instance of moral classes. If lying is forbidden it will be a tuple in beliefset `MoralObligations`. Saying "I'm a millionaire" is a proposition in the second beliefset. This second beliefset contains all states and/or actions for which it is relevant to know whether it is obligatory, permissible or forbidden. These are are evaluated against the first beliefset.

b) Listing beliefset MoralActEvaluation

```
public beliefset MoralActEvaluation extends
OpenWorld {
        #key field String strActName
        #key field String strSphere
        #value field String strActProposition
        #value field String strText
        #value field String strType
        ....
}
```

The above sounds all well and simple. There are some problematic aspects. First, how is an action classified? Or how can an agent recognize it as being subsumed under particular class of moral obligations? E.g. how does an agent know that hitting someone without any cause is not permitted because it goes against the moral obligation 'not to hurt a fellow human being'. Give these questions some thought and it will become immediately apparent that it is far from trivial. It involves understanding the structure of an action or sentence, envisioning the direct and indirect consequences of an action, etc. Second, how does an agent know whether an act is morally relevant or significant? Me

285 In this presentation I will skip several more technical details of the design that have no relevance for the moral aspects. Point in case is the reference. The actual implementation will have an unique ID for referencing purposes plus a name. In the presentation I leave out the unique ID as a field because it has no moral implications.

scratching my ear is a morally uninteresting action, but how do I know whether something is morally interesting/relevant? Or how I can program an agent such that it knows what is of interest and what is not? An option would be to train the agent using neural nets technology. This is possible but results will at the current state of technology be very modest and the process long. We have to provide a fairly complete picture of what are morally salient attributes of acts and states. As these are to some extend situationally determined it will be clear that a complete moral classification will be impossible anywhere in the near future.

At the first step towards implementation it is clear that the logical modelling at the general level is not the problem. The problem arises due to either the absence of clear rules of what is (not) morally relevant in which situation, and, how to analyse actions such that they can be subsumed under moral rules. Our formal understanding of moral epistemology is too fuzzy to be implement for a general purpose agent. Humans can rely on their epistemic capabilities to be trained and learn to recognize situations that are morally relevant and subsume them under the appropriate moral rules. In the artificial context the epistemic capabilities are not yet advanced enough.

This does not mean that nothing can be done. What is required is that acts will have to be restricted and strongly typed (and hence classifiable). The basic structure of the reasoning remains but the classification, structuring and understanding of the acts will be exogenous, i.e. determined at design time. And as time and cognitive science progress these design time decisions can be replaced by stronger epistemic capabilities. What does the short term solution look like? It means that for each act at least its object needs to be defined, the consequence of that act and the evaluation by that object of these consequences need to be known. This means extending the above beliefset with these attributes.

c) Listing extended beliefset MoralActEvaluation

```
public beliefset MoralActEvaluation extends
OpenWorld {
     #key field String strActName
     #key field String strSphere
     #value field String strActProposition
     #value field String strText
     #value field String strType
     #value field String strObject
     #value field String strConsequences
     #value field String strObjectEvaluation
     ....
}
```

The only thing that is now still missing is the actual evaluation. When the agent has properly classified the act it still needs to know whether the act is obligatory or not. To this end the `MoralObligations` beliefset needs to be extended with a query that evaluates an act against the moral obligations. After evaluating it returns an indication that it is obligatory or not.

d) Listing beliefset with function

```
public beliefset MoralActEvaluation extends
OpenWorld {
      ....
      #indexed query getAct
            (String strAct, String strType, boolean
bAct);
      #indexed query getConsequence
            (String strConsequence, String strType,
boolean bConsequence);

      #complex query boolean getObligation (String
strAct, String  strConsequence, String strType){
            boolean bAct;
            boolean bConsequence;
            return getAct(String strAct, String
strType, boolean bAct) &&
      getConsequence(String strConsequence, String
strType,        boolean bConsequence);
      }
      #function query getAllObligatoryActs (){
            ...
      }

}
```

What these queries do is first classify both the act and the consequences using the type indication. And then, based on the classification, query whether both act and consequence are morally obligatory, permissible, etc.

3.2. Desire

Contrary to beliefs desires are easier to implement. A desire or goal is a special event, a `BDIGoalEvent`. It represents the goal-directed behaviour of the agent. The desire to behave morally can be expressed at various levels of detail and complexity. In its simplest form it would look as follows.

e) Listing event as trigger for moral behaviour

```
public event BehaveMorally extends Event
{
}
```

This is admittedly a very crude version but would do the job all the same[286]. A desire is not effective until it is turned into an intention and handled by a plan that can turn the desire into action. It can be tuned further to allow for example the intensity of the desire; the domain (sphere) to which it applies; how and by who or what it is instantiated, etc. The below example show the implementation of these extensions.

f) Listing extended event as trigger for moral behaviour

```
public event BehaveMorally extends BDIGoalEvent
{
      int intensity; //integer denoting the
intensity of the desire
      String strSphere; //application domain
      String strSource; //external source, e.g.
father, mother,...

      #posted as ExternalMotivation (int intns,
String sphr, String src)
      {
            intensity = intns;
            strSphere = sphr;
            strSource = src;
      }

      #posted as ReligousConviction (int intns,
String sphr)
//there is no external source, conviction is
internal
      {
            intensity = intns;
            strSphere = sphr;
      }
}
```

3.3. Intention

286 Depending on how the design is made an event can be kept active till there is a plan that can both process he event and successfully terminates.

An agent has one or more plans. A plan is a sequence of atomic actions that an agent can take in response to an event. Committing to a plan, choosing a plan is like forming an intention. Reasoning about plans, i.e. about which intention to form, is done through meta-level plans. An obligation is a type of plan. It is a sequence of action which should be done. There is nothing fundamental to distinguish an obligation from a non-moral plan as far as they are a sequence of actions. Feeding someone poor and blowing my nose are both a sequence of actions. The distinction is added to the sequence in the meta information we attach to them, in the way we reason about them.
Whether an obligation is adhered to depends on how it is implemented. By tying it closely to an event, and giving it high precedence and prominence its execution can be forced. On the other hand it can be left to meta-level considerations. So an implementation can leave the abidance to the obligation open, and create some uncertainty. Particularly important here is that at design time not all the configurations need to be known (particular event information, precedence, etc. can be determined through configuration data which are read only at run time). A plan as a sequence of actions is a sequence of STIT operators. These are the atomic elements of the action logic.

3.4. Pairwise combinations

Above I discussed the basic building blocks. Evidently they make sense for an implementation only if combined to form complex constructs that express moral attitudes, reasoning, etc. As discussed above the operator O(), the existential and universal quantifier and the moral attribute *G*() are implemented as tuples in beliefsets. As far as the structure and implementation are concerned (Bel i Φ), (Bel i *G*(Φ)), (Bel i E(Φ)), and (Bel i A(Φ)) are the same. Hence I will only discuss the basic forms of intention, belief and desire. The deontic aspect can be added without loss of syntactical validity. The basic operators can be combined to construct morally meaningful propositions that can be to represent desirable properties of moral agents. The intention now is to demonstrate how pairwise combinations of operators can be implemented using the software constructs there were introduced above. Consider the following three pairs[287].

i. (Int i Φ) → (Bel i Φ)[288]

---

287 These propositions are used solely for illustration purposes. I do not argue that they are true or desirable properties of artificial agents.

288 Please note that these pairs serve only the purpose of showing how operators can technically be connected. In these basic forms they might or might not make sense, and be or be not, an desirable feature of a rational software agents (and a moral one at that). Proposition i) states that if i intends Φ, i believes Φ will be. There are many

ii. (Des i Φ) → (Int i Φ)
iii. (Des i Φ) → (Bel i Φ)

List 1 basic pairs of modal operators[289]

The A → B proposition can be given two interpretations. In the first B is a necessary condition for A, the "conditional" interpretation. For implementation purposes one can, for example, think of the context() and relevance() methods which act as conditions for plans to be relevant and applicable. The second interpretation is one in which A is a sufficient condition for B. I will stretch this interpretation by giving it a causal meaning, that is A causes B to happen. I will call this the "causal" interpretation[290]. This makes sense in the context of my approach in which, for example, a change in beliefs causes an event (desire) to be triggered. Can the above three pairs, using these two interpretations, be implemented in JACK?

Ad i) Committing to a plan (forming an intention) requires the context() method to succeed. A context statement can contain a logical proposition with reference to a beliefset. In order to succeed the logical variable(s) need to be bound to one or more tuples from that beliefset. So a belief can be a necessary condition for a plan. Of course the beliefs can be there without the intention being formed.

A plan can operate on beliefsets, adding, changing or removing tuples from beliefset. So there can be a causal relationship between an intention (plan) and a belief.

Proposition i can be implemented under both interpretations of the A → B proposition type.

Ad ii) A desire (event) can cause an intention to arise. This is straightforward pairing where the plan is applicable and relevant to the event. If the plan contains a context() method the desire and belief both appear in the antecedent. If the context() method is empty only the event causes the plan to be instantiated. So there is a direct causal relationship between

reasons why i might intend something which does not come true (something more urgent happening after the intention was formed). So the refined form (Int i Φ) → (Bel i E(Φ)), where E is the existential path qualifier makes sense, but (Int i Φ) → (Bel i A(Φ)), where A is the universal path qualifier, does not make sense. As I am only interested in the technical aspects of the connection I will disregard the plausibility of the proposition at this stage.

289 There are eight combinations of two operators. For the purpose of this article the discussion of three of them suffice.

290 The notion of causality is a complex. By this informal use I do not intend to take any stance in the discussions on the nature of causality. The logical use of conditionals does not imply causation. Here I move from conditional to causal to indicate that in the implementation causality is intended.

As an event (desire) can be triggered from within a plan, the plan (intention) can act as a sufficient condition. Again the relationship is straightforward. And proposition ii can be implemented for both interpretations.
Ad iii) An event cannot operate on a belief. A belief is modified either via external mechanisms and sources or via plans. Hence the occurrence of an event can never be a sufficient condition for a change in beliefs.
Events (desires), on the other hand, are sent through an automatic mechanism for posting: `#posted when ( condition )`. The condition contains a reference to the agent's beliefsets. In this way the beliefset functions as a sufficient condition for the sending of the event (the instantiation of the desire).

The discussion above is summarized in table 1 below. The column 'Conditional' indicates that the antecedent cannot take place without the consequent, it is a necessary condition. The 'Causal' column indicates the cases in which the antecedent is a sufficient condition for the consequent.
Please note that the relationships in the table below are possible relationships. That means that they can be constructed as described. But in the model, in an application, there can be overriding relationships which may cause another relationship not to hold true. In the case of conflicting obligations an additional proposition needs to be introduced to detail how to decide the conflict.

Proposition	**Conditional**	**Causal**
i) (Int i Φ) → (Bel i Φ)	□	□
ii) (Des i Φ) → (Int i Φ)	□	□
iii) (Des i Φ) → (Bel i Φ)	□	x

Table 1 operator connections

4. Complex propositions – negative moral commands

If i believes something to be morally obligatory he form the intention to bring that something about.

11. $(Bel\ i\ O(\Phi)) \rightarrow (Int\ i\ \Phi)$ or $(Bel\ i\ O(\Phi)) \rightarrow (Des\ i\ \Phi) \rightarrow (Int\ i\ \Phi)$

This is a core notion and I will use it as starting point to investigate the implementation of moral notions. Before proceeding there is one further distinction to be made. There are 'obligations to do something' and

'obligations not to do something'. Though this might seem trivial it will become clear that this distinction has substantial implications. The obligation to tell the truth, is not the same as the obligation not to lie.
Let us look at the moral obligation not to kill a fellow human being. How to implement the adherence to the command 'thou shall not kill' ? Rephrase this as killing someone is forbidden, $F\Phi$, or $O\neg\Phi$, where Φ = killing someone. Say we have an agent with the desire to be moral and to adhere to 'do not kill', is expressed in (12)

12. $(\text{Bel } i\ O(\neg\Phi)) \rightarrow (\text{Des } i\ \neg\Phi) \rightarrow (\text{Int } i\ \neg\Phi)$

Implementing this obligation looks as follows. The agent has a beliefset in which the various moral obligations are stored amongst which $\neg\Phi$. I model the beliefset under open world semantics which means that it either holds true, does not hold true or is unknown. This means that it states which moral obligations the agents adheres (not) to. Based on this beliefset it posts a `BDIGoalEvent` for which it seeks applicable plans that help the agent responding to the event.
Now the next question is how to implement obligations? There are two options. Option one, there is a plan Φ representing the obligation. This plan takes precedence, if and when required, over plans to do to contrary. Option two, pre-conditions are added to all plans determining when they are (not) permissible. In this way the obligation cannot be said to have one location, instead it is spread across various plans.
Option one. When a particular desire arises and a `BDIGoalEvent` is raised a set of applicable plans will be selected. Amongst the various plans that are applicable is also plan Φ. As moral obligation it can be given a higher ranking and thereby pre-empt the other courses of action, in particular the ones that would count as violation of the obligation. Does this mean that a plan is permissible if there are no obligations to do the contrary? This is in fact what the formula $P\Phi$, it is permissible that Φ, or $\neg O\neg\Phi$, says. But this seems to me to be problematic. Is shooting someone permissible in the absence of a plan not to shoot someone? What plans does the software agent have 'not to kill'? The answer is none. There is no positive act, no plan to 'not shoot'. Or the set of plans is non-empty and mostly meaningless in the sense that me drinking coffee is not going to kill anyone.
The problem stems from the fact that killing someone is not an act! This is contrary to the way it is usually treated, and the way we talk about it. One cannot define *the* act of killing. This might seem counter-intuitive because we all know what it is to kill someone. Or do we? Try to program an agent to kill someone. The agent would not know what to do directly if told to 'kill someone'. And it is equally problematic, if not more so, to

define what 'not killing' is in a way that can be programmed directly as an act for an agent to execute.
We have many acts that count as killing, e.g. shooting someone, strangling, etc. It is impossible to exhaustively list all acts that count as killing someone or more precise a list of all obligations, not to strangle, not to shoot, etc., etc. This shows that practically speaking from an implementation point of view $P\Phi$, it is permissible that Φ, $\neg O \neg \Phi$, is problematic. 'Thou shall not kill' is a very imprecise statement. It seems to suggest an act but it is in fact about a state. It means 'thou shall not bring about a state of someone not having a beating hart any longer' (or whatever counts medically as being dead). Killing is a reference to a class of acts that bring about the same state. It is a reference to the consequences of an act[291].
In the same vain consider lying. I can breath, I can draw a line, I can smile, but I cannot lie. I can tell you I did not steal the money, where in fact I did steal the money. That would be called lying. But I cannot say “Do (not) lie!” in the same way as “Do (not) raise your hand!”[292]. Try to program an agent 'not to lie'. It does involve a reference to what I bring about (my description of a situation) that does not match what I believe to be the case. So again it is a class of acts with a description that contains a reference to the consequence of those acts. This discussion shows that some obligations cannot be defined as acts. For now I conclude that option one, executing plans in absence of an other, overriding plan that represents a moral obligation, is problematic. The discussion of the implementation shows a different light on moral obligations that is interesting and might open new perspectives.
Option two. Pre-conditions, in the form of the context() method, can be used to control the execution of plans. They can be added, for example, to all plans to kill someone which would then possibly, on evaluation, fail and cause the plan to be excluded from the set of plans up for consideration. These pre-conditions can regulate when a plan should be up for consideration. It will make a reference to the beliefset containing all obligations and possible exceptions, and decide if it counts as a violation.

---

291 The above argument focusses on negative moral commands. The reader might wonder whether the same problems might be relevant to positive moral commands, e.g. to save a life. The reason for not including positive moral commands is that there is an important asymmetry between positive and negative moral commands in relation to acts. Whereas for both the number of acts that count as adhering to is endless in the case of negative moral commands it is important that all are excluded. For positive moral commands it is not important to known them all. In whatever way a life is saved is unimportant as long as it is saved. But with killing we want to make to make that all hundred are identified and stopped rather than ninety-nine, because the one undetected renders are other prevention meaningless.

292 Perhaps I am not pushing this line of argumentation far enough yet. There are obviously many ways in which I can raise my hand. This only emphasises the point I am making: one has to push towards the lowest levels possible in constructing behaviour.

This option does not require a plan that is hard to conceive as in option one. The drawback is that the reasoning about plans is delegated to a lower level. When considering situations in which the act would not count as a breach of an obligation this might prove a problem because the meta-level reasoning is excluded. It also requires particular knowledge at design time about the moral system in which the agents will be functioning and the plan deployed. A plan can be executed in various situations in response to various events (desires). What counts as a valid reason (desire), as valid act (plan) in what circumstances is determined by the moral system under which the whole is executed. Some will be invalidated by all systems, but many will not. On the one hand it is desirable to delegate knowledge at the lowest possible level in the system. On the other hand duplication and inflexibility should be avoided[293].

When the application domain is limited, to say, health care, or even further to sub-domains like hospitals, the number of value changes is limited over a longer time frame, and can be captured relatively easy. Step by step the application domain can be extended without overburdening the design capabilities.

Above I concluded that moral obligations are actually statements about states of affairs that are brought about by particular acts (plans). Acts that are classified based on the outcomes they produce. Following this notion one can combine option one and two and construct plans that have a pre-condition with a reference to the state that the plan may bring about. This variable should be checked against a list of states that may or may not be brought about. All remaining plans for which the deontic status cannot be determined upfront can be dealt with through the various mechanisms of meta-level reasoning. This would be a complete model in the sense that it would catch all plans (intentions) to kill before hand without having to have an exhaustive list before hand. Also the formulation and implementation of states is relatively straightforward. It requires the loading of a list from the configuration file at run-time containing all undesirable outcomes.

The catch with this revised option is that it assumes strong epistemic abilities. Ideally it should be possible for each plan to estimate its impact and consequences. At design time this is at best partially possible. Another part will be dependent on the circumstances that cannot be foreseen at design time. The better the epistemic capabilities the better this options functions. If they absent this option automatically reverts to

293 Korienek and Uzgalis [9] make a very compelling case for redundant degrees of freedom in systems as this increases the adaptability of a system tremendously. The argument is made for artificial life systems, deriving from the study of biological systems. I am convinced that a similar case can be made for software systems as a subset of artificial systems.

option one, because no plans can be excluded upfront. To improve the performance even further plans can be combined into larger combinations of plans that form a capability. The content of each individual plan is reduced, and the 'intelligence' is contained in interaction between these plans. At the higher level plan that integrates the lower level plans the software environment contains a strong function to determine under which conditions a particular goal will be achieved. This function, `@determine`, contains a logical condition and a `BDIGoalEvent`. It finds the conditions under which the goal event can succeed (if at all). This provides an equivalent of 'internal' reasoning before the act to envision the consequences of an act.

The first tentative conclusion is that the mechanisms for the implementation of negative moral commands are available but practical only in limited applications contexts. Moral commands are imprecise. This is their strength, they have a wide domain of applicability. But it is also their weakness: they are vague and open for much debate and interpretation. Negative moral rules are short-cuts for defining classes of acts whose outcomes are undesirable and can be ruled out upfront without further consideration. When the application domain is limited it is practically possible to define such classes, see for example [13]. All the mechanisms are available but require intense computing and/or a better understanding of our moral reasoning.
This leaves a last topic for consideration. What is the role and impact of epistemology? As noted the epistemic requirements can be strong. The last section is left for some initial observations on the role epistemology in the context explicit ethical agents.

5. Epistemology

What has not been discussed above in detail is how an agent comes to belief something, nor where its desires arise from. It will be clear from the above discussions that epistemology plays an important role in morality. In the implementation there are broadly speaking five elements of epistemological nature: 1) knowing the general moral propositions (the commands, rules, etc.); 2) knowledge of the actual state of affairs and the intended states and acts; 3) the projections of the consequence (future states) resulting from particular acts; 4) classification of acts and states under moral rules; 5) perception of moral attributes.
The importance of perception and cognition can be easily recognized. When considering the perlocutionary use of words (what the speaker intends to do by uttering them) and the moral implications, the question is how does an artificial agent distinguish the perlocutionary force of an argument. If goodness is said to be a non-natural characteristic that is

supervenient on the non-evaluative characteristics of a situation then how does an artificial agent perceive or deduce it? If ones duty is discerned intuitively how does an artificial agent discern? What exactly is it that the agent discerns? If an agent beliefs the consequences of some act to be harmful to someone else, and hence refrains from executing this act, how did it come to hold this belief.
These are all questions that point to the epistemic capabilities of an artificial agent. None of these are easy questions. Answering them is outside the scope of this article. But without moral epistemology there can be no fully functional artificial, moral agent, unless one is willing to discard moral theories that rely heavily on epistemic capabilities. One reason for doing so could be the impossibility of implementing these capabilities, and hence the unrealistic nature of these capabilities. At first glance this surely is very unsatisfactory. It might have some merits but insufficient to make such a claim right away. For now we could, perhaps just have to, accept that only a limited set of moral philosophies can be supported in artificial agent environments. Or that the application domain is limited such that the various restrictions, plans, etc. are known at design time and can be strongly typed[294].
One other key question is whether artificial agents have a need for the same type of morality as humans. I think it is desirable to provide artificial agents with some kind of morality, or rule abidance capability (relevant in any complex situation which cannot be modelled completely by an exhaustive set of rules). It seems to me at the current state of development any artificial construct is still limited in its capabilities. It subsequently has no need for a complex moral reasoning capability as humans do, yet.

6. Conclusions

From the research it can be concluded that combined versions of modal logic suffice to model many relevant elements from the normative ethical discourse. Reasoning and meta-level reasoning is supported in sufficient detail to allow for moral reasoning about obligations. Moral theories make strong epistemic claims. The epistemic requirements are key for an implementation of moral reasoning in artificial agents. The current state of both technology and understanding of moral reasoning allow for limited

294 There is a parallel development in speech recognition software. In the early stage of the speech recognition technology the first applications of voice recognition were in specific, clearly demarcated domains like law and medicine. Because of the limited application domain ambiguities in the interpretation of a word could be ruled out upfront, because in the limited context it could only have one meaning. Only after the advancement of technology, a.o. Faster processing of large amounts of data, could the application domain be widened.

implementation only. Full ethical agents are out of our realm. Explicit ethical agents within a limited application domain seem to be possible, though there remains a lot of work to be done in both engineering and moral philosophy. Work at the design and implementation level rather than the fundamentals. Short term research will focus on the full implementation of the constructs presented in this paper.

Acknowledgements

I am grateful to Gert-Jan Lokhorst, Jim Moor and Sabine Roesner for their comments on earlier drafts of this paper.

References

[1] Agent Oriented Software Pty. Ltd, JACK, www.agent-software.com.au

[2] Bedau, M.A., Philosophical content and method of artifical life in The Digital Phoenix, Bynum et. al. (eds), 1998

[3] Bratman, M.E., *Intention, Plans and Practical Reasoning*, Harvard University Press, Cambridge, 1987

[4] J. Broersen, M. Dastani, Z. Huang, J. Hulstijn, and L. van der Torre, The BOID architecture. In the *proceedings of the fifth international conference on Autonomous Agents* (Agents2001), Montreal, 2001

[5] Danielson, P., Artificial Morality, Routledge, London, 1992

[6] Floridi

[7] Governatori, A Formal Approach to Negotiating Agents Development, in *Electronic Commerce Research and Applications*, *1 no. 2*, 2002

[8] Hoven, M.J. van den, Lokhorst, G.J., Deontic Logic and Computer Supported Computer Ethics in *Cyberphilosophy*, Bynum et. al. (eds), 2002

[9] Korienek, G., Uzgalis, W., Adaptable Robots in *Cyberphilosophy*, Bynum et. al. (eds), 2002

[10] Moor, J.H., The nature, importance, and difficulty of machine ethics *(IEEE forthcoming)*, 2006

[11] Russell, S., Norvig, P. (2003) Artificial Intelligence, 2nd edition, Prentice Hall.

[12] Sergot, M., Richards, F., On the Representation of Action and Agency in the Theory of Normative Positions, in *Fundam. Inform. 48 (2-3): 273-293*, 2001

[13] Wiegel, V. Hoven, J. van den, Lokhorst, G-J., Privacy, deontic epistemic action logic and software agents, in *Ethics of new information technology*, proceedings CEPE2005, CTIT workshop proceedings, Enschede, 2005

[14] Wiegel, V., SophoLab *(forthcoming)*, in Ethics and Information Technology, 2006

[15] Wooldridge, M., *Reasoning about Rational Agents*, MIT Press, Cambridge, 2000

COMPUTATION AND REPRESENTATION OF MEANING IN A MAN-MACHINE DIALOGUE

A PRAGMATIC COMBINATION OF LOGICAL FORMALISMS

JEANNE VILLANEAU

Université de Bretagne Sud

VALORIA

OLIVIER RIDOUX

Université de Rennes 1

IRISA

Abstract: We propose an approach to man-machine dialogue in spontaneous spoken language which combines several logic formalisms like illocutionary logic, categorial grammars, conceptual graphs, and formal concept analysis. Our goal is to tackle the difficulties of natural language understanding while keeping the possibility of a generic and extensible system. All this seems to contradict the fact that natural language dialogue does not exactly follow formal rules. In fact, an informal ingredient, a heuristic, is introduced to glue the logic formalisms together and to control their application. For instance, the heuristic allows skipping words, ignoring word order, or naming a whole by one of its parts. A part of this project is concerned with the extraction of the meaning of natural language utterances using a logic-based formalism. This has been implemented as a computer program connected to a speech recognition system. Categorial grammars are used for shallow syntactic analysis, and conceptual graphs are used to formalize ontologies that are used to disambiguate the input. Experiments show that the system is good at recognizing speech hesitations, false-starts, and repairs. Another part, not yet implemented, deals with man-machine dialogue. The plan is to use a formalization of relevance inspired from formal concept analysis in order to maximize it.

8. Introduction

Natural language understanding and dialogue managing are two essential modules in spoken language man-machine systems. They are closely related because the semantic representation of the first one provides the conditions for dialogue management. Currently, human-machine dialogue systems are generally designed for database querying with a very restricted application domain: the state of the art is presented in section 2. In order to extend the understanding capacities of the speech understanding systems, we propose a logical approach of speech understanding which combines syntactic and semantic tools: it is presented in section 3. We have implemented an understanding system according to this approach which is described in section 4. We then present how we want to use formal concept analysis and logical information systems in order to implement dialogue managing, and so to complete this logical approach (section 5).

9. The state of the art

Natural language processing has become an important domain with various research fields among which Man-machine dialogue can be counted.

Implementing a system capable of having a conversation about any topic whatsoever is currently a dream. However, man-machine dialogue systems are already open to the general public. They are designed for very specific tasks and pragmatic context plays here an essential part.

Those systems consist of several modules: the first one is a speech recognizer which performs speech to text conversion. Downstream, a module of natural language understanding (NLU) builds a semantic representation of the utterance. This representation is used by the dialogue manager which links the interface with the database and decides on the answers or queries to send to the speaker.

In such systems, natural language understanding is thus an essential element which may be defined as a translation from the natural language of the human speaker to a formal language which the

computer can use. Such a translation is necessarily accompanied by the loss of a part of the meaning. The first problem to solve is the choice of the formal target language. If it is too poor, too large a part of the meaning is lost, and it makes it impossible to implement a sufficiently natural dialogue. On the other hand, if it is too rich, it becomes too difficult to use it for managing man-machine dialogue.

In natural language processing, computing and representing meaning are considered complex tasks. First of all, many linguists have shown the difference between the meaning of the words and the message that the speaker wants to convey. Ambiguities and implicit meanings are always present in natural language. Context knowledge is very important in order to solve them. Moreover, in spontaneous spoken language, there are many repairs, false-starts and hesitations. Finally, in a human-machine dialogue system, the speech recognizer introduces many errors: words are replaced by phonetically-related words that are not semantically-related.

Two approaches are currently used for natural language understanding. The first approach assumes that meaning can be computed from syntax *[Montague, 1974].* If the utterance is syntactically correct, it is possible to build its syntactic structure: its semantic representation can be inferred from this syntactic structure. This approach is seldom used for “real applications” because the axiom according to which semantics can be deduced from syntax is not very realistic: natural language, especially spontaneous speech, does not obey simple syntax. Utterances are not often syntactically correct if classical criteria used for texts are to be retained; this is aggravated by speech recognition errors. Moreover, even when the method is applicable, extracting semantics from syntax generally means obtaining complex semantic representations written in a rich formal language. The semantic representations thus obtained are often too complex to be easily used by the dialogue manager.

The second approach to natural language understanding is based on pragmatic arguments. All currently man-machine dialogue operational systems are designed for specific tasks in very restricted domains: train time table *[Lamel, 2000]*, weather forecasts *[Zue, 2000]*, etc. In those systems, no complete syntactic analysis is needed. One frequently used solution consists in building semantic frames in order to represent all the possible queries. Understanding can thus be reduced to the detection of phrases or keywords which make it possible to fill in the various fields of those frames *[Bruce, 1975].* For instance, in a flight reservation system, the client word sequence is searched for clues on the origin and destination of a desired flight. Such solutions are robust and effective for the tasks for which they are

designed. In these systems the speaker asks the machine for a service, but it is actually the machine that asks the questions though it shows a different linguistic form. However, when the dialogue form goes wrong with these systems, the escape procedure is that the machine explicitly asks the questions that were implicit in the failed dialog. It is not sure that this is sufficient if the dialogue becomes less constrained, or if the application domain becomes a little less narrow *[Allen, 2001].*

On the one hand, the first approach of speech understanding gives very precise tools that work only for syntactically correct inputs. On the other hand, the second approach can deal with almost all syntactic violations, but it constrains dialogue to fixed frames and it is not sufficient to implement natural man-machine dialogue where the human speaker can slightly stray from simply giving answers to questions raised by the system. In short, to improve spoken language understanding is crucial in the development of man-machine dialogue systems, even if those systems are designed for restricted tasks.

10. Our objectives and the basic principles of our proposal

a) Semantic representation

Our goal is to achieve man-machine dialogue in spontaneous spoken natural language where the range of tasks to be performed is known in advance but is more complex than what can be described by a fixed set of frames. More precisely, our proposal is focused on man-machine dialogue for database querying. The approach we have chosen in order to represent utterance meaning is widely inspired by the speech act theory *[Austin, 1962], [Searle, 1970].* We assume that we know what the dialogue purpose is and that the speaker engages in a dialogue really for data-base querying. We are thus in a very simple dialogue context! But, even in such a context, the man-machine dialogue system must understand what the human speaker wants to say for each speech turn. To do so, it is of course necessary to have a precise semantic representation of the objects present in the speaker sentence, but it is not sufficient. It is also necessary to determine the language act which is applied to these objects. In data-base querying dialogues, many speech acts are not present: for instance, promising or threatening does not have any significance for a machine. Nevertheless, there are various possible acts such as (the

test domain is tourism information with possibilities of hotel or train reservation):

1) simple information request: *“I want to know the fees charged”,*
2) refusal: *“I don't want to go here”,*
3) reservation request (which must change the database state),
4) information: *“I have reserved a room in Caumartin hotel...”*.

More precisely, the semantic representation we have chosen is inspired by D. Vanderveken’s *illocutionary logic [Vanderveken, 2001]*: the logical formula that the system provides is the result of composing a *speech act* and a structure which represents application domain objects that we call *object string*. The logical formula relates objects and properties of the application context according to the domain ontology. Thus, the speech act corresponds to Vanderveken’s *illocutionary force* and the object string corresponds to *propositional content*. The formula is a conceptual graph *[Sowa, 1992]*. It is composed of concepts and conceptual relations that are initially attached to lexical entities in a dictionary, and that the parser attaches to all parts-of-speech, from the most elementary to the whole sentence.

b) Basic parsing principles

The consequence of *hesitations* and *repairs* in spoken language is that the meaning carrier becomes a discontinuous subsequence of the original word sequence. A parser must discover the discontinuous subsequence. A dual point of view is that a discontinuous subsequence of the original word sequence does not carry the meaning; it is *noise*. The man-machine dialogue system has to parse un-syntactic utterances but spoken language studies have shown *[Blanche-Benveniste, 1990], [Martinie, 2001]* that minimal syntactic structures are generally preserved in the repairs and false-starts. Moreover, the meaning of a word association is generally more strongly asserted than the meaning of an isolated word. We have thus chosen to carry out an incremental bottom-up parsing, where words are gradually combined. At the beginning, the parser groups words according to syntactic rules only but, as word groups increase, their meaning becomes more specific and it is thus possible to relax syntactic criteria thereby overcoming the problem of ungrammatical sentences. At the end, when sentences carry complex meanings, they can be composed on semantic grounds, completely disregarding syntactic constraints. The words that participate in this construction form the useful support for the sentence meaning.

The possible links between objects, and between objects and properties, are described in a domain-related semantic knowledge (an

ontology) which expresses in a kind of type system how objects and properties can be compounded. It is of course domain dependent but, in order to preserve the genericness of the other components of the system, it is defined through generic predicates. Thus, the ontology turns out to be the only attachment to a particular application domain. In terms of conceptual graphs, this semantic knowledge can be viewed as a collection of elementary conceptual graphs.

11. An implementation of these principles: the LOGUS system

a) LOGUS description

The LOGUS system implements the principles we have previously presented: incremental and bottom-up parsing where syntactic and semantic criteria are combined with progressive relaxation of the syntactic constraints *[Villaneau, 2004]*.
In a first step, the parsing builds chunks: in those minimal syntactic structures, grammatical words are linked to the lexical words to which they are referred *[Abney, 1991]*. A meaning is also attached to each chunk by consulting the type of its lexical word in a lexicon. The definition we have given for the chunks in LOGUS is more restricted than the classical definition of chunk that are usually used for the parsing of texts: our chunks can contain only one lexical word. This is because forming a chunk is a kind of commitment to the assumed meaning of its components. However, we do not use semantic knowledge at this time because the speech recognition system introduces many recognition errors. The parser, therefore, must not commit too fast. The formation of larger chunks is delayed until it is confirmed by the semantics.
The logical formalism we have used here is categorial grammars *[Bar-Hillel, 1964]* augmented with a semantic component under the shape of typed lambda-terms. Being lexicalized, categorial grammars are well-suited for the analysis of discontinuous subsequences of spoken utterances.

In the following step, rewriting rules are used in order to link the chunks. Those rules are based on both syntactic and semantic criteria and they use the predicates which define the domain semantic knowledge. This step is split into three sub-steps: in the first one, only obvious links are built. In the second sub-step, links are built if classical syntactic and semantic constraints are met and, in the last one, syntactic constraints are relaxed.

In fact, the basic rule of categorial grammars,

if *x* has type *A* and *y* has type *A\B*,
then *xy* has type *B*,

can be relaxed in

if *x* has syntactic type *A* and semantic type α,
and *y* has syntactic type *A\B* and semantic type α□β
then *x noise y* has syntactic type *B* and semantic type β,

and even in

if *x* has syntactic type *A* and semantic type α,
and *y* has syntactic type *A\B* and semantic type α□β
then *y noise x* has syntactic type *B* and semantic type β.

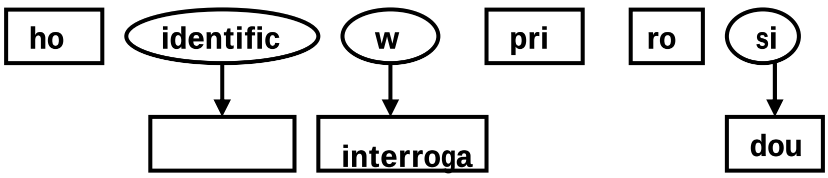
In the original categorial grammar formalism, notation A\B represents a topological constraint on the input. In the semantic component, α□β represents only a logical constraint on the input. The heuristic consists mainly in starting with topological constraints and then relaxing them with logical constraints.

b) An example:

The example presented below shows the succession of the various stages during the parsing of the utterance:

"À l'hôtel Caumartin quel est le prix pour un pour une chambre double"
(In Caumartin hotel what is the price for a for a double room)

After the first step, there are six chunks the semantic translation of which is given below. In the repair *"pour un pour une",* the first word group of grammatical words *"pour un"* has been deleted, because it was not connected with a lexical word.

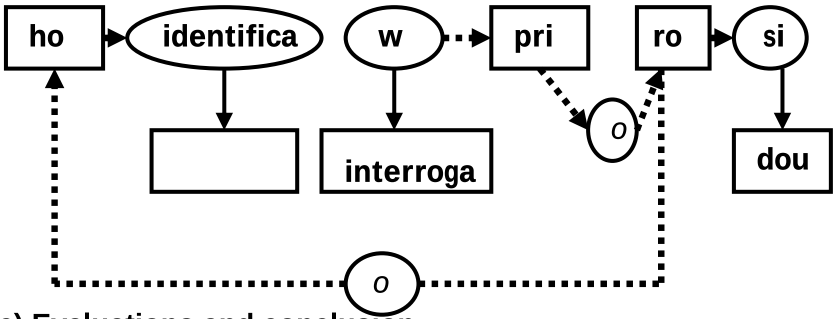


During the second step, links between the chunks are gradually built.

- In the first sub-step, *"hotel"* is linked with *"Caumartin"* because there is a *"Caumartin hotel"* in the domain semantic knowledge.
- Then, links are built between *(what interrogation)* and *(price)*, between *(price)* and *(room)* and between *(room)* and *(size double)*, because all the syntactic and semantic criteria of the rewriting rules are respected.

- At last, the association of chunks *((hotel) (identification Caumartin))* is linked with *(room)* with the subordination conceptual relation *(of)*. Building this link is allowed by the relaxation of a syntactic constraint (word order).

Finally, the logical formula LOGUS provides a parsing result that corresponds to the following conceptual graph. As usually in the conceptual graphs, concepts are in the rectangular boxes and conceptual relations are in the oval boxes.



c) Evaluations and conclusion

To compare the understanding capabilities of various systems is very difficult, because each system has its own semantic representation, related to the choice of the application domain, and the strategy for managing dialogue. Nevertheless, it is an essential task, because one currently does not know which are the most effective approaches to speech understanding implementation.
LOGUS took part in two evaluation campaigns: the DEFI campaign *[Antoine, 2002]* and the MEDIA campaign *[Bonneau-Maynard, 2006].* Those evaluations and our own experiments showed that combining syntax and semantic is an efficient approach in order to implement natural language understanding, with the expected criteria: robustness -when facing repairs and speech recognition errors- combined with precision and capacity to translate correctly quite complex utterances.
We can thus conclude that in spite of imperfections, the LOGUS system answers the conditions that we had set forth. However, in order to prove the interest of our approach, it is necessary to show that the semantic representation provided by LOGUS is efficient when implementing dialogue managing.

12. The dialogue

In a separate work *[Ferré and Ridoux, 2001]*, we have demonstrated that formal concept analysis (FCA, *[Ganter and Wille, 1999]*) can be used as a formal basis to a form of dialogue that is not limited to querying a data-base. In particular, it allows both the human speaker and the machine to ask questions, and it contains a formal notion of relevance, and of its maximization *[Sperber, 2004]*. In fact, a question may be answered by another question, which enables progressive focalization on a definitive answer, and questions are related to the notion of formal concept introduced in FCA. Since formal concepts are ordered in FCA, questions may be ordered; relevance, its maximization and progressiveness may then be formalized. As this vision on the relation between FCA and man-machine dialogue is new, we would like to go into finer details to explain it.

The principle of formal concept analysis is to consider a set of *objects* and then build *formal concepts*. Objects are decorated with properties; this constitutes a *formal context*. Formal concepts (*concept* for short) are sets of objects *O* that share a common property *P*. The essential feature of concepts is that all objects that share *P* are in *O*, and that *P* is the most precise property that all elements of *O* share. The *O* part is called the *extension* of a concept, and the *P* part its *intension*.

Formal concepts can be ordered via their extensions;

a concept *c* is smaller than a concept *c'*, if *extension(c)* is included in *extension(c')*.

In a dual way, concepts can be ordered via their intensions;

a concept *c* is smaller than a concept *c'* if *intension(c)* entails *intension(c')*.

The two orders are equivalent; in both cases, one says that *c* is a *sub-concept* of *c'*.

Some concepts (not all) can be generated by starting from an object *o*, and computing the set of all objects whose property entails the property of *o* (thus, intension is *property(o)*, and extension is *extension(property(o))*). Such an object *o* is a paradigm of the generated concept; it has all properties, and only them, of this context. It is possible that several objects generate the same concept; in some sense these objects are equivalent, they carry the same information.

In a dual way, some concepts (not all) can be generated by starting from a property *p*, and computing the set of all objects whose property entails *p*, and then computing the most precise property common to these objects (thus, extension is *extension(p)*, and intension is *intension(extension(p))*). It is possible that several non-logically

equivalent properties generate the same concept; we say that these properties are *contextually equivalent*, because at least in the considered context they express the same concept.
Finally, it is also possible that two properties *p* and *p'* generate two concepts *c* and *c'* such that *c* is smaller than *c'*, though *p* does not entail *p'*. Then, we say that *p contextually entails p'*. Given *p* that contextually entails *p'*, it is always possible to compute maximal properties *p"*, in the contextual entailment ordering, such that each *p"* expresses the difference between *p* and *p'*; *p"* is called the *increment* of *p* over *p'*.

The conclusion of all this is that FCA permits to combine logical entailment and contextual knowledge in a new form of entailment that contains both. The definition of the contextualized entailment relation is strictly computational. We believe this can answer past objections on the use of entailment to characterize conversational appropriateness, like in Grice's maxims. The general idea is as follows. In a query answering application, the machine knowledge is represented as a formal context; and a user expresses queries as a property of what he is looking for (not necessarily a characterizing property). At this stage, a usual query answering system answers with the extension of the query *[van Rijsbergen, 1986]*.
We suggest instead that the query answering system compute the formal concept generated by the query; then check if objects exist that also generate this concept; and finally check for increments that generate sub-concepts of the query concept. The objects, if any, form the extensional part of the answer. The increments form the intensional part of the answer; it is the main originality of our proposal. They correspond to queries asked by the machine to the user:

> Q: Do, you have objects with property *p*?
> A: I have these objects, o_1, … o_n, that are typical of *p*, and I have other objects of various kinds, p_1, …, p_m. Which kind do you want?
> Q: I prefer p_i.
> …

By construction, the o_i and p_i are relevant to query *p*; furthermore the p_i are maximally relevant.
Our contention is that these operations and a few others can serve as a basis for a form of man-machine dialogue where roles are more symmetrical than when querying usual data-bases. Moreover, the dialogue is more progressive because the user can start with a very small indication of what he wants; it is the machine that will offer the details (the p_i) that will form the complete query. This also respects

the principle which states that it is always easier to recognize a description than to invent one.

So, we suggest that the semantic representation of an utterance be treated as in the formal concept analysis approach. The formal context will be formed by a representation of the world and its ontology.

13. Discussion

Natural language communication is not entirely carried by words. Context also carries meaning. This is the part played by formal context analysis in our proposal. We also know that spontaneous speech tends to preserve the grammaticality of noun phrases *[Blanche-Benveniste, 2002]*. This is the part played by chunk analysis. Finally, it is well known that an utterance can be understood even if its syntax is completely broken. This is the part played by the ontology-based heuristics.

The semantic representation part of our proposal has been implemented and tested. Experiments show that meaning of spontaneous spoken utterances can be retrieved in presence of repairs, hesitation, etc. Failures have been observed when meaning is very difficult to analyze, even for a human mind. However, the human mind knows that it does not understand, while the computer-based system proposes a meaning even when it has not understood anything, and should have asked for explanation. In most human-human dialogues, there are understanding errors; however a majority of dialogues “succeed”. Thus, understanding errors must be regarded as normal incidents. A good dialogue manager must have tools to detect and to treat them. This is still the domain of further work.

To conclude, we propose to use formal linguistic models with heuristics to compute the meaning of natural language utterances. Our objectives are modest, though attainable. We wish to stress that our use of formal models is always generic, thereby giving us the assurance that our objectives can be made more ambitious progressively. Indeed, as opposed to frame-based understanding where no composition law exists -so that one cannot see how to tackle complexity- our approach uses formal components that feature powerful composition laws. Thus we can expect to be able to tackle

complexity in a compositional way. Past experience has shown that each one of these formal components cannot deal with the whole spectrum of natural language understanding ; we propose to make them collaborate, each one in its competence domain.

14. References

[Abney, 1991] Abney S. *Parsing by Chunks.* In Principle Based Parsing. R. Berwick, S. Abney and C. Tenny Editors, Kluwer Academic Publishers.

[Allen, 2001] Allen J., Byron D., Dzirovska M., Ferguson G. Galescu L. and Stent A. *Towards conversational Human-Computer Interaction.* AI Magazine, 22(4) pp. 27-37.

[Antoine, 2002] Antoine J-Y., Bousquet-Vernettes C., Goulian J., Kurdi M-Z., Rosset S., Vigouroux N. and Villaneau J. *Predictive and Objective Evaluation of Speech Understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the french CNRS.* The 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain.

[Austin, 1962] Austin J-L. *How to do thing with words.* Oxford

[Bar-Hillel, 1964] Bar-Hillel Y. *Language and Information.* Chapter in "On Categorial and Phrase Structure Grammars". Adison-Wesley, Reading.

[Blanche-Benveniste, 1990] Blanche-Benveniste C. *Le Français parlé : études grammaticales.* CNRS Editions, Paris.

[Blanche-Benveniste, 2002] Blanche-Benveniste C. *L'étude grammaticale des corpus de langue parlée en français.* 2002]

[Bonneau-Maynard, 2006] Bonneau-Maynard H., Ayache C., Bechet F., Denis A. Khun A., Lefevre F., Mostefa D., Quignard M., Rosset S., Servan C. and Villaneau J. *Results of the French Evalda-Media evaluation campaign for literal understanding.* The 5th Int. Conf. on Language Resources and Evaluation (LREC 2006) pp. 2054-2059, Genoa, Italy.

[Bruce, 1975] Bruce B. *Case Systems for Natural Language.* AI Magazine, vol 6, pp. 293-326.

[Ferré and Ridoux, 2001] Ferré S. and Ridoux O. *Searching for Objects and Properties with Logical Concept Analysis.* Proc. of the Int. Conf. on Conceptual Structures, pp. 187-201, Springer.

[Ganter and Wille, 1999] Ganter, B. and Wille, R. *Formal Concept Analysis*, Springer.

[Lamel, 2000] Lamel L., Bennacef S., Bonneau-Maynard H., Rosset S. and Gauvain J-L. *Recent developments in spoken language systems for information retrieval.* Proc. of the European Conf. on Speech Technology (Eurospeech'95), pp. 207-210.

[Martinie, 2001] *Remarques sur la syntaxe des énoncés réparés en français parlé.* Recherches sur le Français parlé, 16, pp. 189-206.

[Montague, 1974] Montague R. *Formal Philosophy.* Yale University Press, New Haven, USA.

[Searle, 1970] Searle J. R. *Speech Acts.* Cambridge University Press.

[Sowa, 1992] Sowa J. *Conceptual Graphs*, Knowledge-Based Systems, 5(3), pp. 171-172,
(see also http://users/bestweb.net/sowa/cg/cgstand.htm).

[Sperber, 2004] Wilson D. and Sperber, D. *Relevance Theory*. In G. Ward and L. Horn (eds.) Handbook of Pragmatics. Oxford: Blackwell, pp. 607-632.

[Vanderveken, 2001] *Essays in Speech Act Theory.* Chapter in Universal Grammar and Speech Act Theory, pp. 25-62. John Benjamin, Amsterdam Philadelphie, D. Vanderveken and Susumu Kubo (eds.).

[van Rijsbergen, 1986] van Rijsbergen C.J. *A new theoretical framework for information retrieval*, Int. ACM SIGIR Conf. on Research and Development in Information Retrieval.

[Villaneau, 2004] Villaneau J., Antoine J-Y. and Ridoux O. *Logical Approach to Natural Language Understanding in a Spoken Dialogue System*.Text, 7th Int. Conf. on Speech and Dialogue, (TSD)., LNAI 3206, pp. 637-644, Brno, Czech Republic.

[Zue, 2000] Zue V., Seneff S., Glass J., Polifroni J., Pao C., Hazen T. and Hetherington L. *JUPITER: Telephone-Based Conversational Interface for Weather Information.* IEEE Transactions on Speech and Audio Processing, 8(1) pp. 100-112

computers & philosophy

i-C&P 2006

Conference Web Site:

http://ateliers.iut-laval.univ-lemans.fr/i-CaP_2006//